Escola Superior d'Arxivística i Gestió de Documents
Adscrita a la UAB

LA VIGILANCIA Y EL CONTROL DE LA POBLACIÓN A TRAVÉS DE LA GESTIÓN, LA CONSERVACIÓN Y LA EXPLOTACIÓN DE DATOS MASIVOS

Autora: Eva Mejias Alonso

Director del trabajo: Joan Soler Jiménez

Curso 2015/2016

Máster en Archivística y Gestión Documental

Escola Superior d'Arxivística i Gestió de Documents

Colección: Trabajos fin de máster y de postgrado

Cómo citar este artículo: Mejias Alonso, Eva. (2016) *La vigilancia y el control de la población a través de la gestión, la conservación y la explotación de datos masivos*. Trabajo de investigación del Máster de Archivística y Gestión de Documentos de l'Escola Superior d'Arxivística i Gestió de Documents. (Trabajos fin de Máster y de postgrado). Http://... (Consultado el...)

Resumen

Este trabajo analiza cómo es la gestión documental en un data center desde una óptica archivística, teniendo en cuenta los sistemas y tecnologías que intervienen en la recopilación, gestión, conservación y análisis de datos masivos.

Así mismo, reflexiona acerca de las consecuencias que ello supone en la privacidad de los datos y refleja cómo la gestión documental puede ser, en este caso, una potente herramienta de vigilancia masiva.

Para ello, toma como estudio de caso los data centers de algunas de las agencias de inteligencia de Estados Unidos y de la Unión Europea y los data centers de las principales empresas de Internet y telecomunicaciones.

Palabras clave: [Big Data, Vigilancia masiva, Data center, Gestión documental, Agencias de inteligencia, NSA, Datos masivos, Centro de datos, Control social, Bases de datos]

Title: Population monitoring through the management, conservation and exploitation of massive data

Abstract

This dissertation examines the archival perspective of how record management is developed in a data center, focusing on how the system and technology are involved in massive data collection, management, conservation and analysis.

This paper reflects how record management can be a powerful tool of mass surveillance and also, the consequences of non-privacy.

To develop and delve into the theories, we have taken as a case study some American and European intelligence agencies' data centers and major Internet and telecommunication company's data centers.

Keywords: [Big Data, Mass surveillance, Data center, Record management, Intelligence agencies, NSA, Massive data, Social control, Data bases]

Sumario

Índice de figuras

1 Introducción

El presente trabajo pretende dar a conocer el tipo de gestión documental que se realiza en un data center, analizando desde una perspectiva archivística la gestión, la conservación y la explotación de datos masivos. Así mismo, partiendo del hecho de que existen miles de estudios sobre las bondades del Big Data, el trabajo pretende mostrar sus consecuencias sobre la privacidad y las libertades ciudadanas.

La investigación toma como estudio de caso los data centers de las agencias de inteligencia de las principales potencias democráticas, acotando la zona de influencia a Estados Unidos y la Unión Europea.

Partiendo del hecho de que gran parte de los datos que recopilan estas agencias proceden de grandes empresas privadas, también se analizaran los data centers de las principales compañías de Internet y telecomunicaciones.

Para ello, se estudiaran las tecnologías empleadas en los diferentes procesos de la gestión documental y los distintos medios de análisis de datos masivos que permiten a las organizaciones implicadas obtener resultados concretos que vulneran múltiples aspectos éticos y legales.

El trabajo se estructura alrededor de cuatro apartados:

- **Conceptos básicos:** un apartado introductorio fundamental para comprender dos de los conceptos más influyentes del estudio: Big Data y data center. Se definen ambos conceptos y se explican sus características principales.

- **Los data centers de Silicon Valley:** determina y justifica las empresas privadas cuyos data centers serán objeto de investigación. Analiza las técnicas y tecnologías empleadas en la recolección, almacenamiento y análisis de los datos, así como el uso que hacen de ellos.

- **Los data centers de la vigilancia masiva:** es el grueso del trabajo. Determina y justifica los gobiernos y las agencias de inteligencia que se estudiarán y se centra en la recolección, el almacenamiento, el análisis y la explotación de los datos masivos en los data centers de las agencias de inteligencia.

- **Legislación:** repasa la evolución de las principales leyes, normativas y recomendaciones que se han llevado a cabo desde las filtraciones de Edward Snowden en el 2013. Es un apartado que ayuda a comprender cómo han reaccionado y actuado ante el escándalo de la vigilancia masiva los órganos de poder público de los Estados Unidos y de la Unión Europea.

1.1 Objetivos del trabajo e hipótesis

Objetivos

- Analizar los procesos de gestión documental que intervienen en un data center.
- Estudiar las tecnologías utilizadas en los procesos de gestión documental de un data center.
- Analizar el uso de los datos masivos y sus consecuencias éticas y legales.
- Reflejar la vigilancia masiva indiscriminada ejercida por gobiernos democráticos a través de los data centers estudiados.

Hipótesis

El trabajo parte de las siguientes hipótesis:

- Un data center puede ser considerado como un archivo.
- La explotación de los datos por parte de las agencias de inteligencia estudiadas en este trabajo busca el control de la sociedad.
- Las tecnologías Big Data vulneran la privacidad de los datos.
- Los gobiernos democráticos mencionados en este trabajo son partícipes del uso de datos por parte de las agencias de inteligencia para ejercer la vigilancia masiva indiscriminada.

1.2 Metodología

La metodología del estudio se ha basado en la revisión de la bibliografía disponible sobre el tema, en la extracción de las ideas y datos pertinentes de la bibliografía consultada y en el posterior análisis y comparación de los datos extraídos.

Las fuentes de información utilizadas son múltiples: artículos de prensa, artículos de revistas especializadas, monografías especializadas, páginas web, blogs especializados, estudios universitarios, tesis doctorales e informes de compañías privadas como Oracle, Microsoft o Telefónica.

La búsqueda bibliográfica parte especialmente de buscadores y bases de datos como CCUC, CRAI, CSIC, Dialnet, Doaj, Google Scholar, Open Doar, RACO, Recercat, etc.

Cabe destacar en este apartado que gran parte de la documentación relacionada con los data centers de las agencias de inteligencia proviene de los documentos que el ex analista de la CIA, Edward Snowden, filtró a la prensa en junio de 2013. Toda la documentación filtrada es accesible a través de la página web edwardsnowden.com[1].

La documentación en cuestión ha sido considerada como auténtica, fiable y válida por múltiples medios de comunicación, por organizaciones internacionales como la ONU y por el Parlamento Europeo.

[1] Journalistic Source Protection Defence Fund. *Snowden doc search* [en línea]. 2013. Disponible en: <https://search.edwardsnowden.com/>.

1.3 Estado de la cuestión

Hoy en día, cuando alguien habla de la Agencia de Seguridad Nacional de los Estados Unidos (NSA), la conversación se centra normalmente en la privacidad, y con razón. Sin embargo, no es el único tema al respecto que vale la pena discutir. Me llama más la atención hablar de la NSA como un estudio de caso a la hora de analizar la ingente cantidad de datos que una organización puede recoger, almacenar y tratar y ver qué consecuencias comporta para la población el fenómeno Big Data.

Big Data es un término omnipresente en la actualidad. Hace referencia a la ingente cantidad de información que proviene de todas partes y que hoy día podemos procesar, analizar y utilizar… para bien o para mal. Es la revolución de los datos, el sistema que ha modificado la manera de hacer negocios y sectores como la sanidad, la política o la educación. Y a su vez, es el responsable del fin de la privacidad de los datos de toda la ciudadanía.

Las agencias de inteligencia siempre han dependido de las comunicaciones a la hora de obtener información. La interceptación en redes electrónicas no es nada nuevo, pues los ingleses ya usaron en los años 40 un sistema de vigilancia electrónica muy compleja para espiar comunicaciones de oficiales alemanes. Lo que es nuevo es el volumen y el alcance de la información que hoy en día pueden interceptar.

Si bien es cierto que las filtraciones de Edward Snowden en junio de 2013 marcaron un antes y un después a la hora de hacernos una idea del alcance y de las capacidades de ciertas agencias de inteligencia, años antes la NSA ya estaba construyendo en Utah un centro de datos de proporciones gigantescas con capacidad de albergar un número incierto de exabits procedentes de sus actividades de vigilancia.

Es inquietante pensar que en plena crisis económica los gobiernos de Estados Unidos y de la Unión Europea hayan invertido billones de euros procedentes de los bolsillos de los ciudadanos en financiar tecnología de espionaje masivo. La cuestión del uso de Big Data por parte de las agencias es que todo el mundo es un objetivo, porque Big Data trata datos que provienen de infinidad de fuentes. No hay discriminación. Puede que algunas agencias de inteligencia y grandes empresas privadas como Google tengan unas herramientas de análisis de datos masivos extraordinarias, pero recogen tal cantidad de información que ni las tecnologías Big Data más avanzadas pueden evitar que la mayoría sea irrelevante de cara a sus objetivos.

Si supuestamente el objetivo principal de agencias de inteligencia como la NSA es la seguridad nacional, la inmensa mayoría de datos que reciben de comunicaciones y transacciones de ciudadanos de todo el mundo son inútiles, porque no están relacionadas con el terrorismo, el espionaje o con cualquier otro tipo de crimen.

Se ha demostrado que los datos masivos no son necesariamente la mejor fuente de información a la hora de resolver ciertas cuestiones, tal y como hemos podido ver estos últimos años, donde no se ha conseguido ni prever ni evitar ninguno de los atentados que se han producido en Estados Unidos, Bélgica o Francia.

Faltan datos de calidad, y eso hace pensar que tal vez valdría más la pena estudiar otros métodos de recolección y escrutinio donde las fuentes de información sean realmente relevantes. En este ámbito que nos es aún desconocido a la mayoría de archiveros, podría ser este el punto de inflexión de nuestra profesión.

El hecho de que nuestro colectivo se encuentre tan alejado de la gestión de la información en ámbitos tan tecnológicos como un data center (o centro de datos) es el principal motivo de ser del trabajo. La información y la sociedad cambian y evolucionan a una velocidad vertiginosa, y no estamos adaptándonos a ella. Cada vez son más las empresas y organizaciones que requieren de data centers para gestionar grandes volúmenes de datos, y en pocos años la tendencia se incrementará considerablemente, por lo que debemos reivindicar la utilidad que nuestra profesión puede aportar en este sector.

2 Conceptos básicos

2.1 Big Data

2.1.1 Definición

El concepto de Big Data, o datos masivos, es relativamente reciente y desde hace varios años está muy presente en todo tipo de ámbitos, por lo que no hay una definición única al respecto, ya que dependerá del sector o de la especialidad de quien haga uso de esta tecnología o sistema. A modo de ejemplo, a continuación podemos ver una serie de definiciones formuladas por distintos sectores que, aunque en esencia expresan lo mismo, incorporan matices diferentes.

En el artículo de Mario Tascón para la revista Telos, *Introducción: Big Data. Pasado, presente y futuro*[2], podemos ver que el autor realiza una definición bastante extensa sobre el concepto. Telos es una revista especializada en tecnologías de la información, por lo que se trata de una definición hecha des del ámbito de las TIC:

*"Big Data es todo aquello que tiene que ver con grandes **volúmenes** de información que se mueven y analizan a alta **velocidad** y que pueden presentar una compleja **variabilidad** en cuanto a estructura y composición […] También es importante comprender que además de los datos estructurados, aquellos otros que provienen de fuentes de información conocidas y que, por tanto, son fáciles de medir y analizar a través de los sistemas tradicionales, empezamos a poder y querer manejar datos no estructurados: los que llegan de la web, de las cámaras de móviles y vídeos, redes sociales, sensores de ciudades y edificios…"*

[2] Tascón, Mario. Introducción: Big Data. Pasado, presente y futuro. *Telos: Revista de Pensamiento sobre Comunicación, Tecnología y Sociedad* [en línea]. Julio-septiembre de 2013, núm.95 [Consulta: 17 junio 2016]. Disponible en: <http://telos.fundaciontelefonica.com/seccion=1268&idioma=es_ES&id=2013062110090002&activo=6.do>.

Por otro lado, en el artículo *Big Data: una "revolución industrial" en la gestión de los datos digitales,* de la empresa Fidelity Worldwide Investment[3], dedicada a proporcionar servicios de gestión a activos e inversores, se muestra una definición mucho más breve y concisa que la anterior:

"Big Data es el término en inglés que designa conjuntos de datos de gran tamaño y generalmente desestructurados que resultan difíciles de manejar usando aplicaciones de bases de datos convencionales".

Ahora bien, si nos fijamos en las definiciones de Big Data que dan grandes gigantes de la informática, como Oracle e IBM, en sus respectivos artículos *Big Data y su impacto en el negocio*[4] y *Analytics: el uso de big data en el mundo real*[5], podemos observar que son muy distintas entre sí.

Por un lado, en el artículo de Oracle se define el concepto de Big Data de la siguiente manera:

"Proliferación de información en crecimiento acelerado y sin visos de ralentizarse, explosión de indicadores, señales y registros, interacciones en redes sociales… […] El mundo está lleno de señales, signos, datos y piezas de información que analizadas y puestas en relación podrían responder a cuestiones que nunca hubiéramos imaginado poder preguntar […] El conjunto de toda esta explosión de información recibe el nombre de Big Data y, por extensión, así también se denomina al conjunto de herramientas, técnicas y sistemas destinados a extraer todo su valor".

Por otro lado, el artículo de IBM lo define así:

"… *Big data es una combinación de las uves –* ***velocidad, volumen y variedad*** *– que crea una oportunidad para que las empresas puedan obtener una ventaja competitiva en el actual mercado digitalizado. Permite a las empresas transformar la forma en la que interactúan con sus clientes y les prestan servicio, y posibilita la transformación de las mismas e incluso de sectores enteros.*"

---

[3] Fidelity Worldwide Investment. *Big data: una "revolución industrial" en la gestión de los datos digitales* [en línea]. Fidelity Worldwide Investment, 2012, p.1 [Consulta: 17 febrero 2016]. Disponible en: <https://www.fondosfidelity.es/static/pdfs/informes-fondos/Fidelity_ArgInvSXXI_BigData_Sept12_ES.pdf>.

[4] García Huerta, Ana. *Big Data y su impacto en el negocio: Una aproximación al valor que el análisis extremo de datos aporta a las organizaciones* [en línea]. Madrid: Oracle, 2012, p.6 [Consulta: 17 febrero 2016]. Disponible en: <https://emeapressoffice.oracle.com/imagelibrary/downloadMedia.ashx?MediaDetailsID=2197>.

[5] IBM Institute for Business Value. *Analytics: el uso de big data en el mundo real : Cómo las empresas más innovadoras extraen valor de datos inciertos* [en línea]. IBM Global Business Services, 2012 [Consulta: 17 febrero 2016]. Disponible en: <http://www-05.ibm.com/services/es/bcs/pdf/Big_Data_ES.PDF>.

Mientras Oracle destaca especialmente el conjunto de datos desestructurados y las herramientas de extracción de valor de datos, IBM se centra más en las denominadas tres uves, que se explicaran más adelante, y define el término desde un punto de vista enfocado al sector empresarial.

Acercándonos más al mundo de la gestión documental, la documentalista Aina Giones-Valls define Big Data en su artículo de la revista BiD, *Cuantificarse para vivir a través de los datos: los datos masivos (Big data) aplicados al ámbito personal*[6], aportando una cuarta uve que no aportan el resto de definiciones: la veracidad.
"*Se consideran datos masivos cuando el volumen, la veracidad, la velocidad y la variedad de los datos (las cuatro V de los datos masivos) superan la capacidad establecida por los mecanismos tradicionales de capturar, gestionar y procesar datos en un tiempo razonable*".

Así pues, ¿cómo podríamos definir el concepto de Big Data de una manera clara y que reúna las características dadas en las definiciones anteriores?
A raíz de las principales ideas de dichas definiciones y de varias más dadas por diversos especialistas en la materia, se extrae la siguiente definición que unifica algunos de los conceptos que se mencionan en ellas:

Big Data designa a todos aquellos conjuntos de datos de gran tamaño que no pueden ser capturados, almacenados ni analizados con el software y la infraestructura tradicionales que se han empleado hasta ahora, así como al conjunto de herramientas, técnicas y sistemas destinados a extraer el valor de estos datos para enriquecer y complementar sistemas con capacidades predictivas.

A esta definición, que recoge las ideas clave que se mencionan en los diversos estudios al respecto, se le han de sumar además varios elementos imprescindibles que conforman Big Data, como los **datos desestructurados** y las **tres uves**.

---

[6] Giones-Valls, Aina. Cuantificarse para vivir a través de los datos: los datos masivos (big data) aplicados al ámbito personal. *BiD, textos universitaris de Biblioteconomia i Documentació* [en línea]. Junio de 2015, núm. 34 [Consulta: 17 junio 2016]. Disponible en: <http://bid.ub.edu/es/34/giones.htm>.

Los datos desestructurados

En primer lugar, es necesario diferenciar los datos estructurados de los no estructurados.

Los datos estructurados son aquellos que están almacenados, clasificados y organizados en bases de datos. Se gestionan y analizan en base a una serie de procedimientos concretos, con atributos y mediante indexación.

Por su parte, los datos no estructurados son aquellos que provienen de canales menos tradicionales: redes sociales, mails, fotografías, vídeos, sensores, blogs, servicios de geolocalización…

Los datos no estructurados tienen una gran utilidad predictiva, pero darles el tratamiento que se usa con los datos estructurados sería extremadamente costoso y en ocasiones imposible. Sin embargo, los sistemas Big Data, gracias a una serie de requerimientos técnicos, son capaces de extraer, almacenar, organizar y analizar este tipo de datos.

Las tres uves

En segundo lugar, la ingente cantidad de datos que puede procesar Big Data es sólo uno de sus aspectos. El volumen de los datos a almacenar es básico, pero hay otros atributos igual de importantes, y aquí es donde entran en juego las tres uves, que acaban de definir las características de Big Data: volumen, variedad y velocidad.

Volumen	Cantidad de datos que se pueden almacenar, que se incrementan cada año y que se crean tanto dentro de las organizaciones como a través de la web, dispositivos móviles, sensores…
Variedad	Tipos de datos que se tratan, pues pueden ser datos no estructurados, semiestructurados - como los provenientes del social media -, o los que provienen de información basada en la localización.
Velocidad	Velocidad a la que se crean nuevos datos y la necesidad de analizarlos en tiempo real para extraer valor de ellos. Esta velocidad se incrementa gracias a la inmediatez de las transacciones, a la informática móvil y al creciente número de usuarios de Internet y de dispositivos móviles.

Muchos estudios y expertos sobre Big Data coinciden además en señalar una cuarta uve: el **valor**, ya que el valor económico de los datos varía significativamente dependiendo del tipo de información. Siempre hay información valiosa oculta entre las grandes cantidades de datos no tradicionales, y el reto está en identificar aquellos datos que sí tienen valor para extraerlos y analizarlos.

Además del valor, hemos podido ver en la definición de Aina Giones-Valls otra cuarta uve: la **veracidad**, ya que la incertidumbre no deja de estar presente en los datos que se recopilan a través de las técnicas de Big Data. Es decir, los datos que provienen de sensores automáticos o de técnicas Big Data como el análisis de sentimientos pueden no ser completamente fiables.

Y tal y como muestra el informe de 2015 de la OBS Business School[7] sobre Big Data, a las tres uves convencionales se le han de sumar cuatro más: las dos uves anteriormente mencionadas, que son el valor y la veracidad, más la variabilidad y la visualización.

2.1.2 Funcionamiento

¿De dónde proceden los datos?

Ésta es la primera cuestión, saber de qué fuentes se extraen los datos. En primer lugar, cabe decir que más del 80% de los datos que se generan en el mundo están desestructurados y crecen 15 veces más rápido que los estructurados.

Todos los días se escriben comentarios en Facebook y en Twitter y se suben vídeos a YouTube, pero las redes sociales son sólo uno de los muchos catalizadores de Big Data. Los sensores conectados en red recogen ingentes cantidades de datos de los teléfonos móviles, de los contadores del gas y la luz, de equipos atmosféricos... Los satélites registran datos meteorológicos y geográficos e información para uso militar, se crean datos a partir de cualquier actividad cotidiana y se almacenan datos de transacciones, como los que recogen las cajas de los supermercados.

[7] OBS Business School. *Big Data 2015* [en línea]. Barcelona: Universitat de Barcelona, 2015 [Consulta: 4 mayo 2016]. Disponible en: <http://www.obs-edu.com/es/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet>.

De forma general, Big Data recoge datos prácticamente de todas partes, no solamente del medio digital: redes y medios sociales, dispositivos móviles, transacciones vía Internet, sensores y dispositivos de redes, consultas y resultados de los motores de búsqueda, datos meteorológicos y astronómicos, vigilancia militar, datos económicos y bursátiles, historiales médicos, experimentos físicos, archivos fotográficos, radio y televisión, Internet de las cosas…

¿Qué tecnologías utilizan los sistemas Big Data?

Existe una gran variedad de técnicas y de tecnologías procedentes de diversos campos, entre ellos la estadística, las ciencias computacionales o las matemáticas, que se han desarrollado y adaptado para extraer, analizar y visualizar datos por parte de Big Data.

La tecnología que se utiliza en los sistemas Big Data está constantemente en proceso de desarrollo y mejora. Big Data, básicamente, se beneficia de hardware estándar que permite implantar técnicas de procesamiento paralelo y escalabilidad, emplea almacenamiento no relacional para procesar datos no estructurados y semi estructurados y aplica tecnologías avanzadas de análisis y visualización de datos para extraer elementos de comprensión a los usuarios finales. Relacionado con esto, las tecnologías que destacan especialmente son Hadoop, el lenguaje NoSQL y el procesamiento masivo en paralelo (MPP).

Hadoop[8] es probablemente la tecnología para Big Data más conocida. Es un marco de software de código abierto gratuito que permite trabajar con miles de nodos[9] y petabytes, e incluso exabytes, de datos. Es un proyecto de Apache y funciona mediante el lenguaje de programación Java. En lugar de procesar un enorme bloque de información cada vez, Hadoop segmenta Big Data en múltiples partes, de modo que pueden procesarse y ser analizadas al mismo tiempo. Un cliente puede acceder a datos no estructurados y semiestructurados desde distintas fuentes.

[8] Apache Software Fundation. *Hadoop* [en línea]. 2014. Última actualización: 11 febrero 2016. [Consulta: 20 marzo 2016]. Disponible en: <http://hadoop.apache.org/>.

[9] En el ámbito de la informática y de la telecomunicación un nodo es un punto de intersección, unión o conexión entre varios elementos que confluyen en el mismo lugar. En estructuras de datos dinámicas un nodo es un registro que contiene un dato de interés y al menos un puntero para referenciar a otro nodo.

Las aplicaciones de Hadoop en Big Data son diversas: permite obtener valor de los datos de una organización mediante las tres famosas uves explicadas con anterioridad, permite extraer los datos tanto estructurados como no estructurados de grandes conjuntos de datos, proporciona lenguajes de alto nivel y despliegues automatizados en la nube…

Yahoo está siendo de momento el mayor contribuyente y hace un uso constante de Hadoop en su negocio, y a día de hoy, varios proveedores como IBM, Cloudera o Amazon, ya han integrado Hadoop en sus soluciones propietarias para reducir el coste final de sus productos. Por su parte, cada vez más compañías utilizan esta y otras opciones de software libre para experimentar cómo almacenar, gestionar y analizar grandes cantidades de datos.

La tecnología NoSQL, a diferencia del lenguaje SQL que es un lenguaje de consulta estructurado que interpreta los datos almacenados en tablas y esquemas de bases de datos relacionales, analiza los datos en origen. Gracias al lenguaje NoSQL se obtiene una ventaja de tiempo real sobre las tecnologías actuales y permite procesar grandes cantidades de datos desestructurados. Google y Amazon son dos empresas que utilizan este tipo de lenguaje, ya que analizan páginas web y documentos buscando una palabra clave en lugar de consultar una base de datos relacional centralizada. Combinando bases de datos NoSQL con Hadoop se pueden descubrir patrones prácticamente en tiempo real.

Finalmente, el procesamiento masivo en paralelo (MPP) permite utilizar muchos procesadores informáticos funcionando en paralelo para analizar datos. Anteriormente eran grandes superordenadores los que realizaban esta tarea. Hadoop utiliza este tipo de tecnología y es lo que le ha hecho tan popular, ya que gracias al uso que hace Hadoop del MPP las empresas pequeñas pueden utilizar sus redes de ordenadores de oficina para analizar datos complejos a un coste relativamente reducido.

¿Cómo se visualizan los datos?

La visualización de los grandes conjuntos de datos procesados por los sistemas Big Data es fundamental para poder comprender lo que se ha analizado y procesado y tomar medidas al respecto. Gracias a la visualización se presentan de la forma más comprensible posible los resultados del análisis previo de datos que ha sido llevado a cabo. Las aplicaciones de visualización más comunes y básicas son las *tag clouds*, los *clustergrams*, los flujos historiales y los flujos de información espacial.

Tag clouds: la clásica nube de etiquetas. Es un listado visual donde las palabras más grandes son las que aparecen con mayor frecuencia y las más pequeñas las que aparecen con menor frecuencia. Es un tipo de visualización que ayuda a que el analista pueda percibir rápidamente los conceptos más sobresalientes de un gran cuerpo de texto.

Tabla 1. Tag cloud. (McKinsey Global Institute, 2011)

Clustergram: se utiliza para analizar clústeres[10]. Lo que muestra es cómo se asignan los miembros individuales de un conjunto de datos a un clúster a medida que el número de clústeres se incrementa. Gracias a este tipo de gráficos el analista puede comprender cómo varían los resultados de una determinada agrupación en función del nombre de clústeres.

[10] Un clúster es una colección de objetos que son "similares" entre sí. Un conjunto de clústeres es la división de los datos en grupos de objetos similares, un proceso que ayuda a la agrupación estructural de un conjunto de datos.

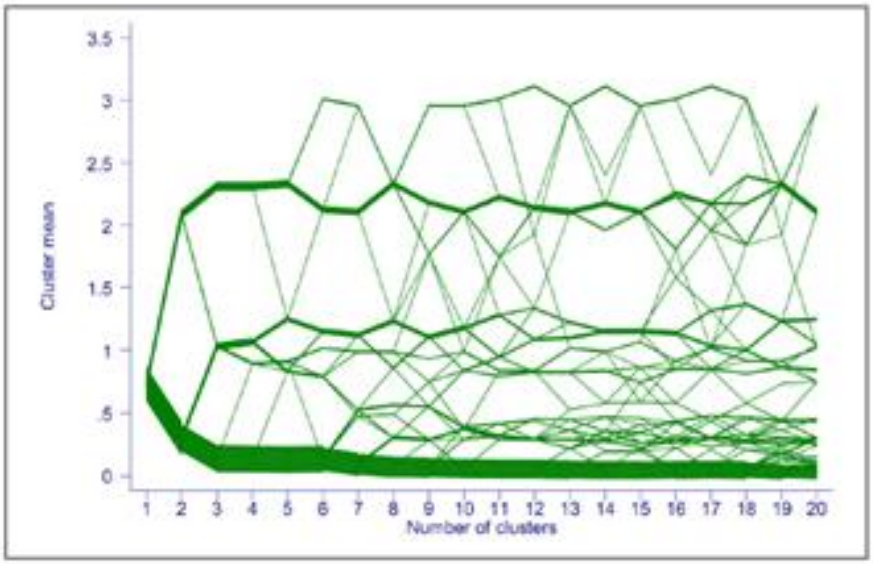

Tabla 2. Clustergram. (McKinsey Global Institute, 2011)

Flujos historiales: trazan la evolución de un documento a medida que éste va siendo modificado por diversos autores contribuyentes. El eje horizontal representa el tiempo y el eje vertical las contribuciones. Cada autor tiene un código de color, y la longitud vertical de cada barra indica la cantidad de texto escrito por cada autor. En la imagen se muestra la entrada “Islam” de la Wikipedia:

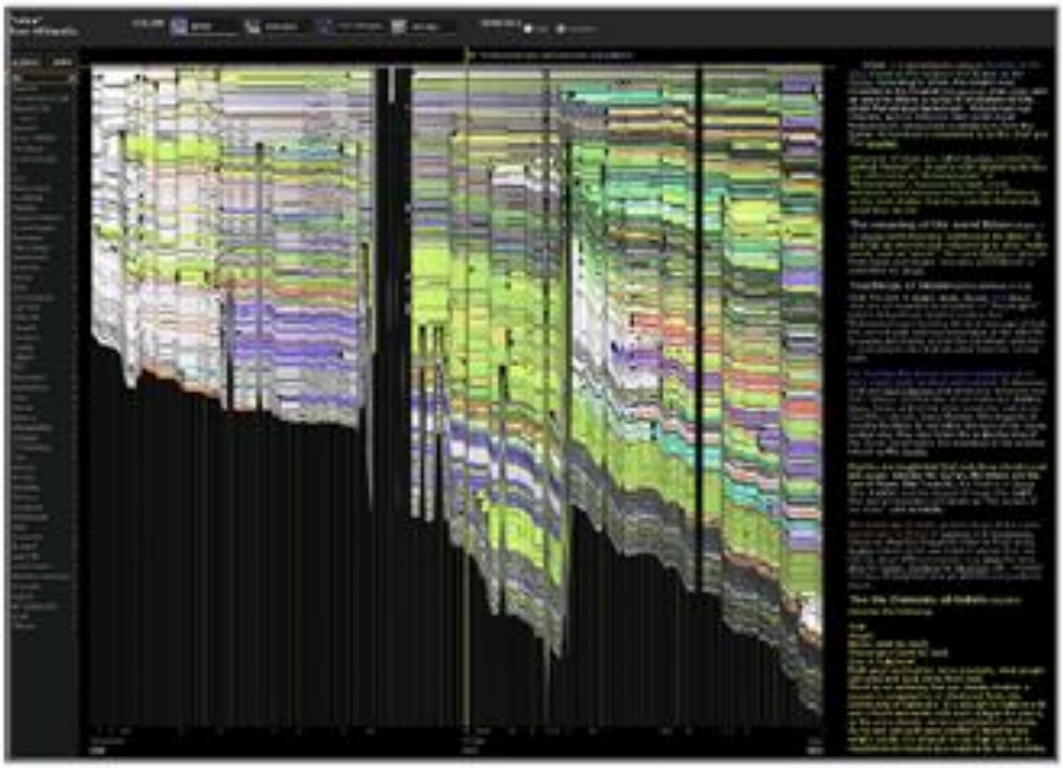

Tabla 3. Flujo historial. (McKinsey Global Institute, 2011)

El gráfico muestra que un creciente número de autores han hecho contribuciones a la historia de esta entrada, así como que la longitud del documento ha crecido con el tiempo a medida que más autores han ido entrando más información. Cuando la longitud de las barras disminuye indica que ha habido eliminaciones de información importantes o que incluso se ha llegado a eliminar el documento por completo.

Flujos de información espacial: como su nombre indica, son imágenes que representan flujos de datos espaciales. En la siguiente imagen se muestra la cantidad de datos de IPs que fluyen entre Nueva York y ciudades de todo el mundo. El tamaño de la luz en una localización particular de la ciudad corresponde a la cantidad de tráfico IP que fluye entre ese lugar y la ciudad de Nueva York. Cuanto mayor sea el brillo mayor será el flujo. Esta visualización permite determinar en este caso qué ciudades están más estrechamente vinculadas a Nueva York según su volumen de comunicaciones.

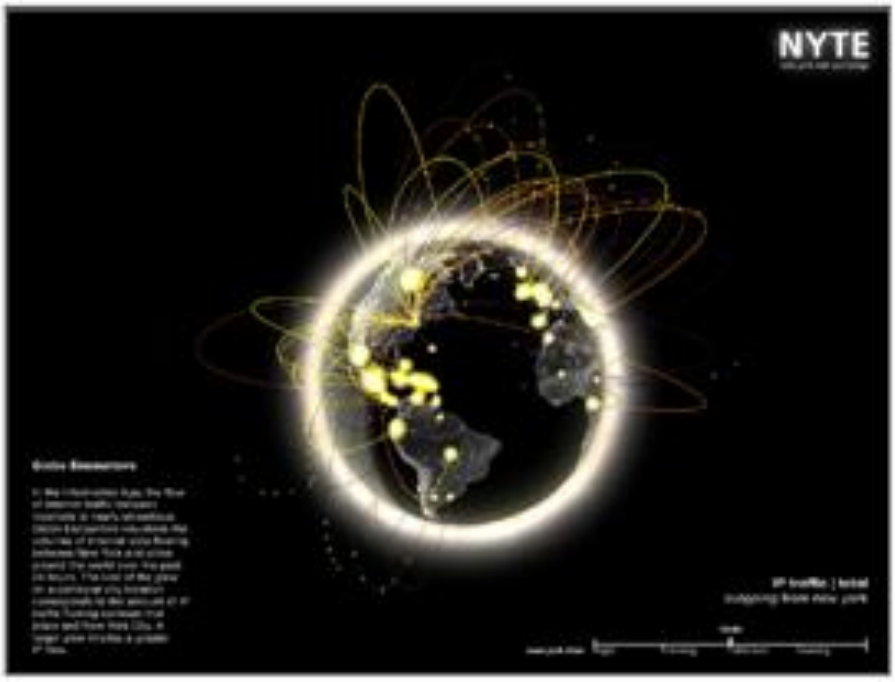

Tabla 4. Flujo de información espacial. (McKinsey Global Institute, 2011)

Infografías: representación visual que resume o explica la esencia de una serie de conjuntos de datos. Son un recurso muy popular en la actualidad. Recoge los resultados de diferentes análisis sobre datos y los presenta de manera simplificada y gráfica, algo que resulta atractivo para un gran número de audiencias.

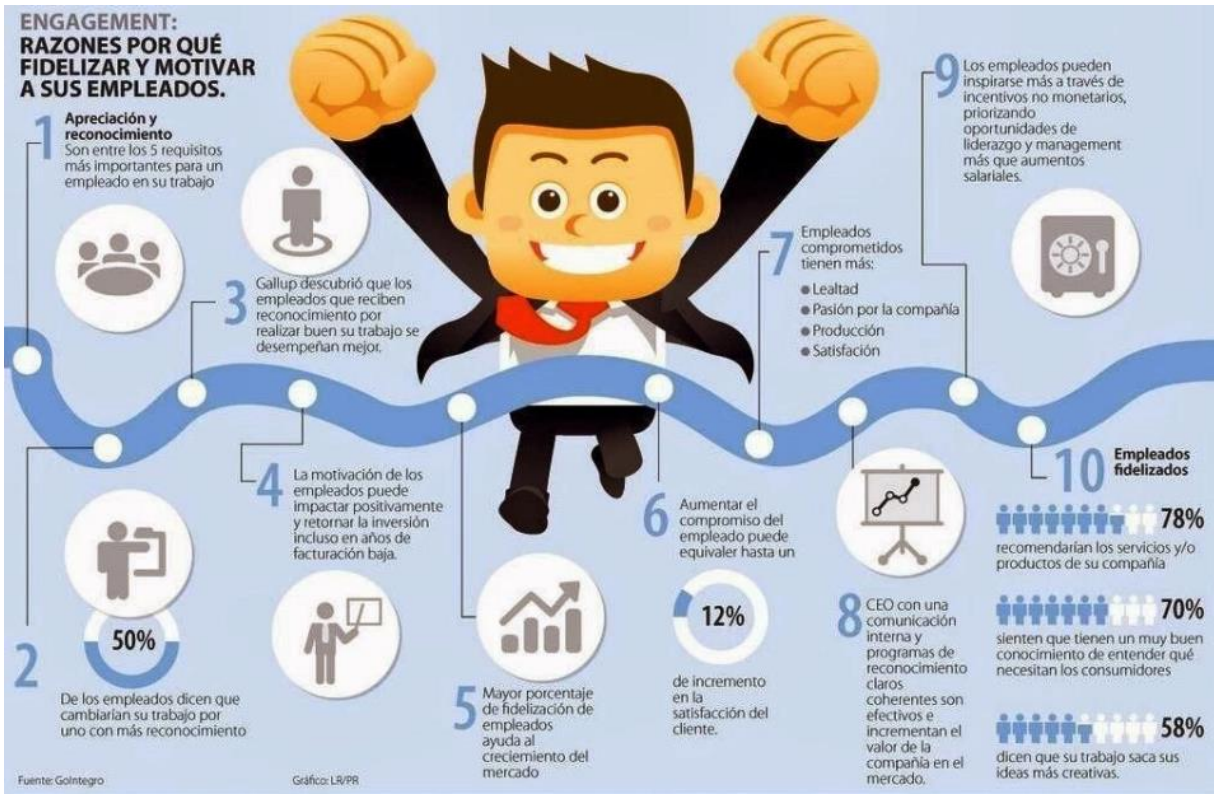

Tabla 5. Ejemplo de infografía. (Olberg Sanz, 2013)

Además de estas técnicas comunes, el mercado se ha llenado de empresas que ofrecen soluciones muy prácticas, dinámicas e interactivas a la hora de visualizar grandes conjuntos de datos.

Un ejemplo de ello lo tenemos en Tableau Software[11], una de las empresas líderes en la visualización de datos que ofrece soluciones rápidas y sencillas para crear visualizaciones de todo tipo de datos e información para poder detectar tendencias, identificar oportunidades y tomar decisiones con seguridad.

En la tabla siguiente, podemos ver de una manera muy visual el estado del tráfico en Singapur. Cada columna del gráfico y cada punto del mapa proporcionan información extra, y además, se actualiza a tiempo real.

[11] Tableau Software. *Tableau* [en línea] 2013-2016 [Consulta: 5 febrero 2016]. Disponible en: <http://www.tableau.com/es-es>.

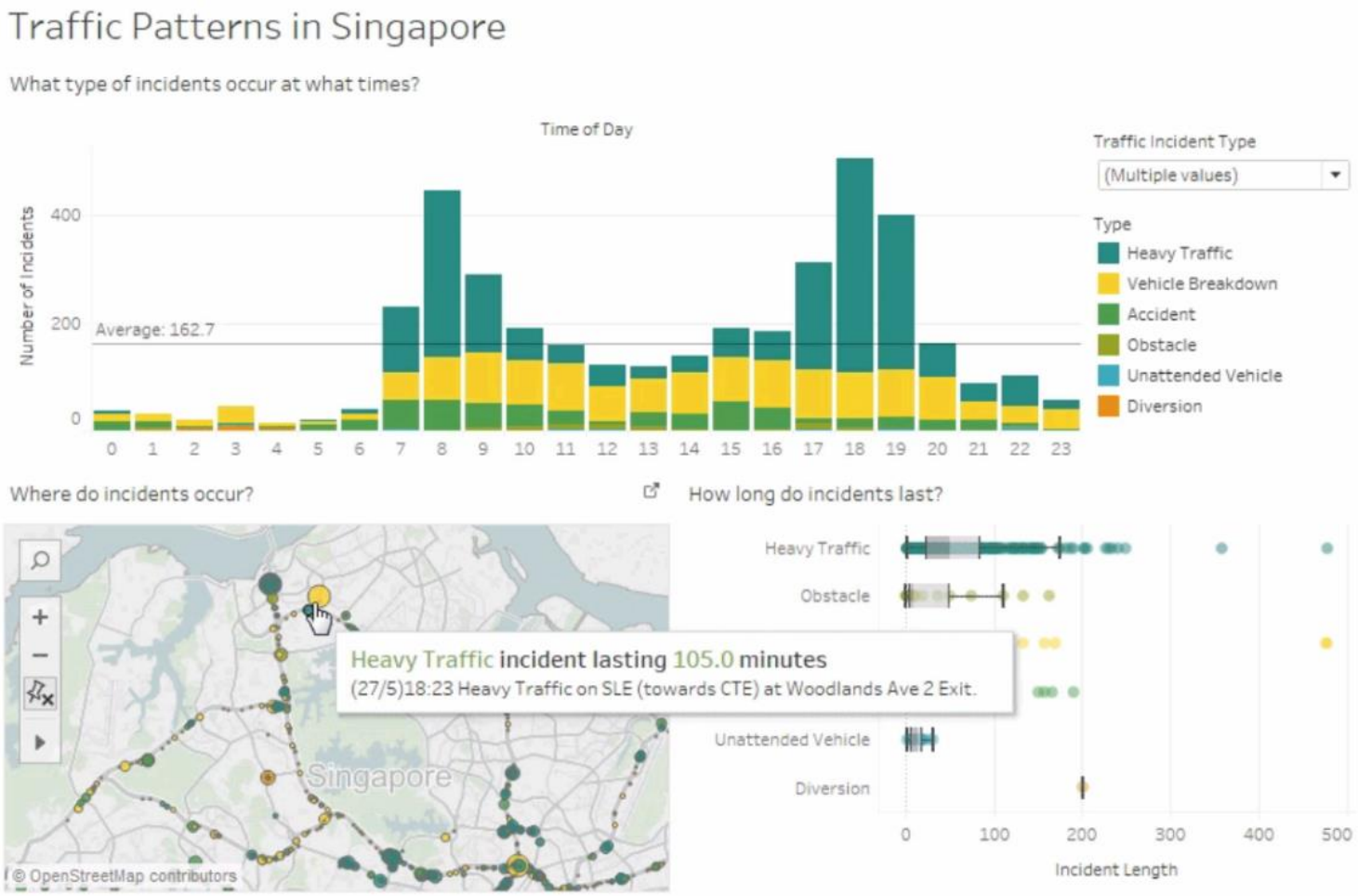

Tabla 6. Patrones del tráfico en Singapur. (Tableau Software, 2016)

Otro ejemplo parecido lo encontramos en Qlik[12], una plataforma de análisis visual. Al igual que Tableau Software, se caracteriza por ofrecer herramientas intuitivas y fáciles de utilizar.

En la siguiente tabla se puede observar cómo se visualiza uno de sus productos diseñados para empresas, pensado para generar informes y cuadros dinámicos y hallar patrones y oportunidades entre los datos que se gestionan.

[12] Qlik Tech International AB. *Qlik* [en línea]. 1993-2016 [Consulta: 5 febrero 2016]. Disponible en: <http://global.qlik.com/es>.

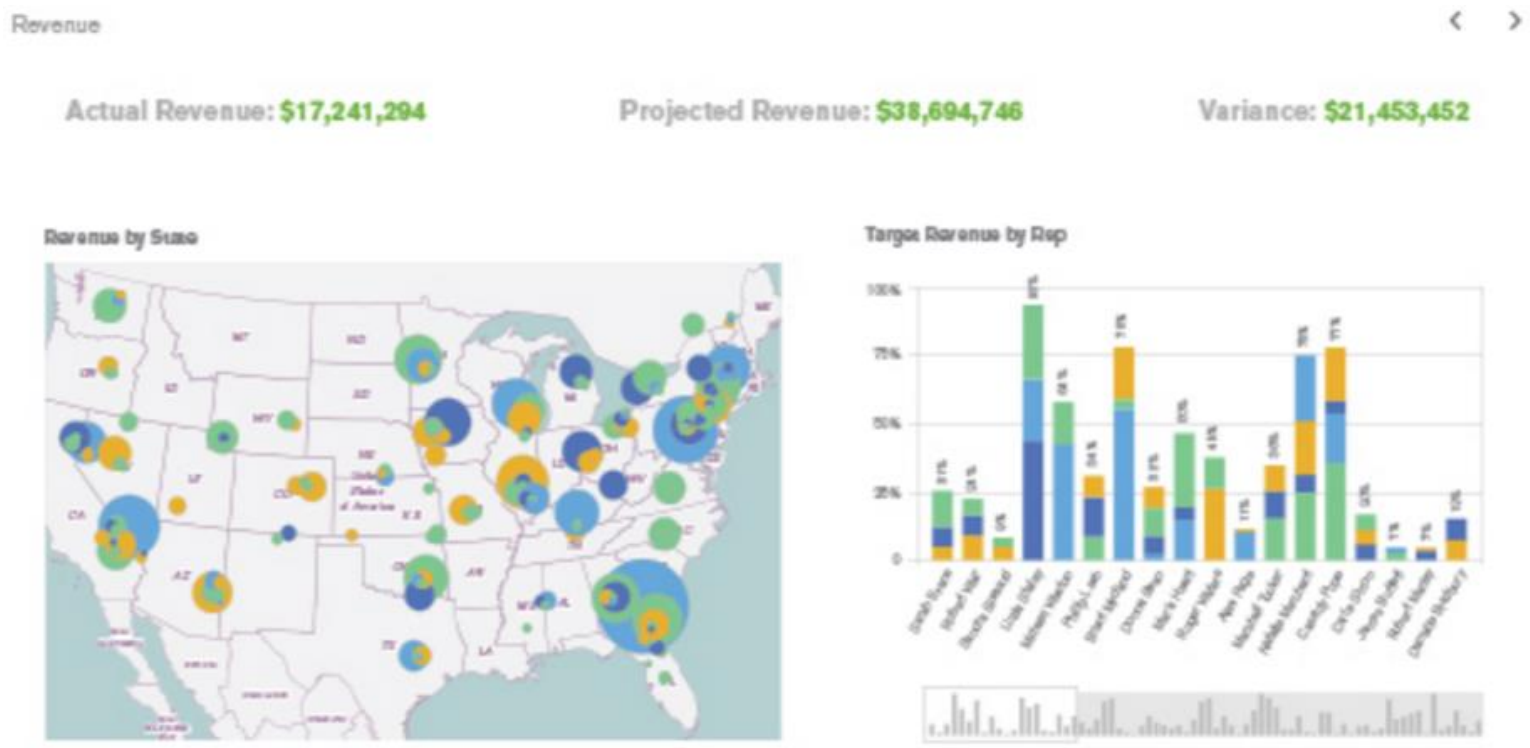

Tabla 7. Visualización de los datos en Qlik Sense. (Qlik, 2016)

2.1.3 Utilidades

El análisis inteligente de datos masivos tiene un gran valor económico para todo tipo de instituciones y organizaciones de prácticamente todos los ámbitos. Sus efectos sobre la economía mundial cada vez son mayores. En España, por ejemplo, se estima que las tecnologías Big Data están teniendo un crecimiento anual en torno al 30% y el 40%, y además, el 62% de las empresas españolas aplican técnicas para el uso de datos masivos.

Según el informe sobre Big Data en 2015[13] de la OBS Business School, la inversión mundial en sistemas e infraestructura Big Data rondaba los 116.000 millones de euros en 2014. Por sectores, durante ese mismo año se registró un aumento de las inversiones en todos los sectores empresariales, siendo Medios y Comunicación el principal inversor con el 53% de las organizaciones del segmento.

[13] OBS Business School. *Big Data 2015* [en línea]. Barcelona: Universitat de Barcelona, 2015 [Consulta: 4 mayo 2016]. Disponible en: <http://www.obs-edu.com/es/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet>.

Si se analiza únicamente a Europa, los sectores económicos más beneficiados por el Big Data son Comercio (47.000 millones de euros), Industria (45.000 millones), Administración Pública (27.000 millones) y Sector Sanitario (10.000 millones).

En su siguiente informe de 2016[14] destaca que, por tercer año consecutivo, Big Data es el principal destino de las inversiones mundiales, la principal fuente de empleo cualificado y la mayor causa de creación de empresas de servicios relacionadas con los sistemas de información. En los próximos dos años, la OBS calcula que aproximadamente el 75% de las empresas globales invertirán en sistemas Big Data.

Cualitativamente, las áreas de negocio donde el Big Data ha demostrado ser un gran aliado de las empresas son en la mejora de la experiencia de cliente y en la mejora de la eficiencia de los procesos de negocio, seguido de transporte, la salud, los medios de comunicación, los seguros, la banca, las comunicaciones y el comercio minorista.

Si nos centramos en los distintos sectores de negocio se pueden concretar las siguientes utilidades:

- **Banca y finanzas:** servicios de protección de marca, protección ante riesgos y fraude y servicios personalizados a clientes.
- **Sector público:** servicios de inteligencia, defensa y protección (control e interceptación de las comunicaciones, vigilancia masiva, interceptación de redes de telefonía e Internet), gobiernos abiertos, open data y proyectos de Smart Cities.
- **Sanidad:** monitorización remota de pacientes, localización de emergencias, almacenamiento de historias clínicas, radiografías, escáneres…
- **Gran consumo al por menor:** control de la cadena de fabricación, análisis de los tickets de la compra, marketing personalizado, fidelización de clientes.
- **Turismo:** optimización de precios y generación de ofertas personalizadas.
- **Telecomunicaciones:** control de la red, venta de servicios de localización, servicios de publicidad asociados al patrón de las llamadas o de las aplicaciones descargadas, obtención de perfiles enriquecidos de consumidores, segmentación y personalización de ofertas, análisis de abandono.

[14] OBS Business School. *Big Data 2016* [en línea]. Barcelona: Universitat de Barcelona, 2016 [Consulta: 4 mayo 2016]. Disponible en: <http://www.obs-edu.com/es/noticias/estudio-obs/estudio-obs-big-data-2016>.

- **Servicios públicos:** uso de contadores y sensores inteligentes, control de la red de comunicaciones, de tuberías, del metro…
- **Web y Digital Media:** análisis de clicks, streamings[15], personalización, forecasting[16] y optimización.

Eso sí, hay que tener en cuenta que Big Data, además de poder ser el activo más decisivo en una organización, también puede ser una de sus obligaciones más costosas. Si se es incapaz de gestionar las herramientas Big Data supone una considerable pérdida de visibilidad de oportunidades y amenazas, el incumplimiento de normativas y pérdidas en las ventas y en la atención al cliente.

Big Data es tan importante para las empresas precisamente por la relación que hay entre el análisis de datos y los resultados de negocio. Las empresas que gestionan mejor sus datos toman mejores decisiones y obtienen mejores resultados financieros.

Hoy en día, Big Data se perfila como la principal fuente de ventajas competitivas para todos los sectores. Las organizaciones y los fabricantes de tecnología que enfoquen esta práctica como una moda pasajera se exponen a quedarse atrás y pronto se verán a sí mismos imitando a rivales con mayor capacidad de pensar un paso más allá.

[15] Técnica de distribución de multimedia donde el usuario consume el producto mientras éste se descarga.

[16] Estimación de la demanda futura de un producto a través de diferentes técnicas de previsión.

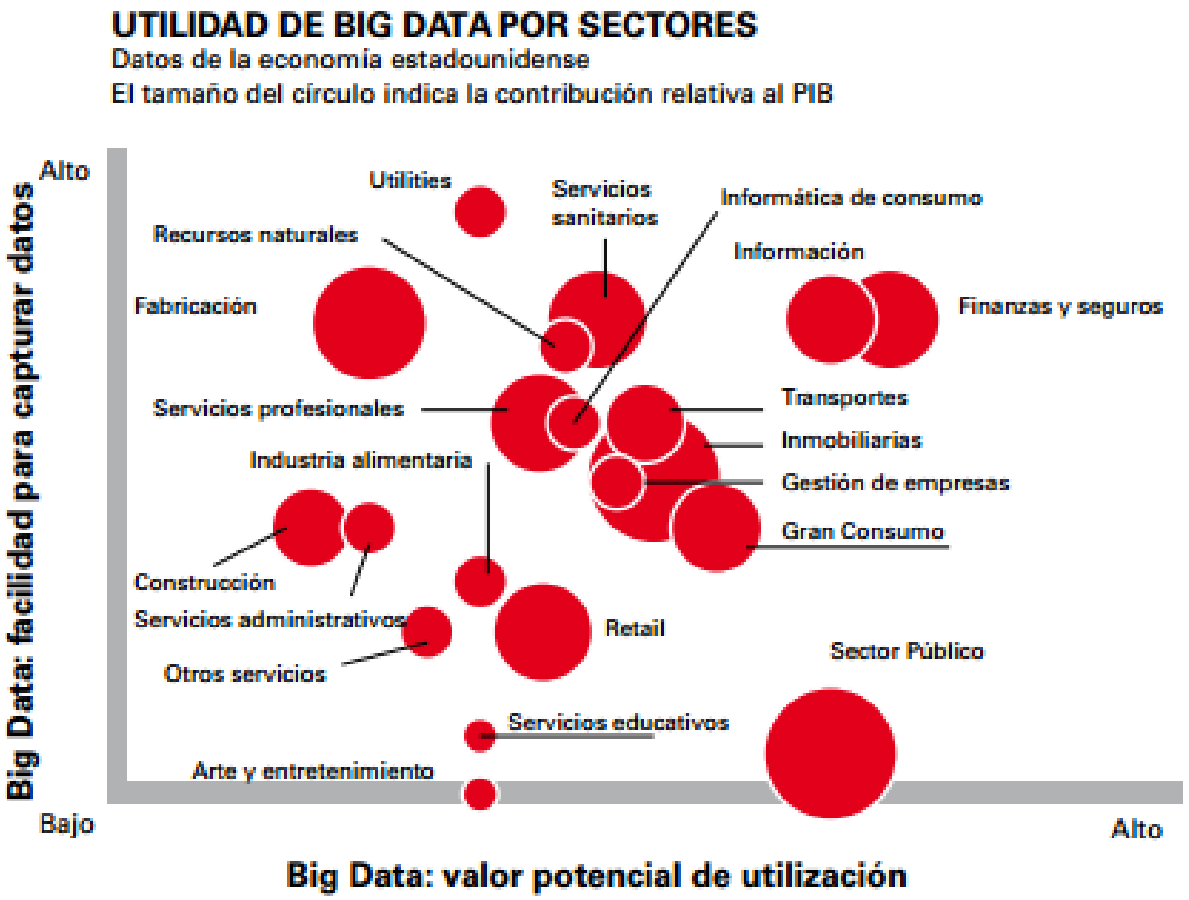

Tabla 8. Utilidad de Big Data por sectores. (Oracle 2014).

Utilidades para los gobiernos

¿Cómo puede usar el gobierno el poder del Big Data para el bien de la sociedad? Big Data no es solamente un medio potentísimo para el control de la ciudadanía, como ya se demostrará en apartados posteriores, sino también un mecanismo capaz de mejorar la calidad de vida.

Si los gobiernos son capaces de aprovechar las oportunidades que ofrecen los sistemas Big Data podrían diseñar políticas públicas que mejorasen notablemente áreas como la movilidad, la salud o la seguridad ciudadana.

A modo de ejemplo, destacan en este sentido algunos casos en Estados Unidos que a través de tecnologías Big Data se han desarrollado diversas mejoras y facilidades para su sociedad:

Predicción de terremotos antes de que sean detectados: en agosto de 2012 el Servicio Geológico de Estados Unidos rastreó miles de tuits buscando la palabra "terremoto". Utilizando datos relacionados con la hora y la posición, consiguió localizar un gran terremoto en las islas Filipinas antes de que los sismógrafos lo registraran.

Alerta temprana sobre focos de epidemias: Google descubrió en 2009 que existía una estrecha relación entre el número de personas que realizaba búsquedas relacionadas con la gripe y las personas que realmente sufrían síntomas gripales. Actualmente, esta herramienta conocida como Google Flu Trends, ofrece datos sobre la actividad de la gripe en diferentes países y regiones de todo el mundo.

Anticipación de la tasa de desempleo: Global Pulse, una iniciativa de Naciones Unidas para abordar crisis socioeconómicas en colaboración con la empresa SAS, analizó conversaciones que tenían lugar en las redes sociales de Estados Unidos e Irlanda y consiguió predecir aumentos en la tasa de desempleo tres meses antes que los informes oficiales.

Prevención de delitos y reducción de índices de criminalidad: si bien es cierto que las tecnologías Big Data han potenciado la eficacia a la hora de combatir el crimen, destaca especialmente el controvertido caso de la ciudad de Los Ángeles, donde los sistemas Big Data se usan para detectar qué grupos o individuos son más propensos a cometer delitos y deben ser sometidos a una vigilancia extra.

Medicina 4P (personalización, predicción, prevención y participación): Big Data puede ayudar a entender cómo funcionan los genes y cómo interactúan entre ellos para prevenir y luchar contra enfermedades como el cáncer. Permite analizar los datos genéticos de cada individuo, y por lo tanto personalizar los tratamientos médicos. Gracias a Big Data pueden estudiarse efectos secundarios perjudiciales de medicaciones o interacciones entre distintos medicamentos, y utilizar esta información para configurar tratamientos a medida que optimicen los resultados y minimicen los riesgos. Hoy ya es posible prevenir ataques al corazón o comas diabéticos analizando los datos generados por dispositivos de monitorización de uso doméstico, como medidores de glucosa o monitores cardíacos.

2.2 Data Centers

El otro concepto clave en este trabajo es el data center, ya que el principal objetivo de esta investigación es estudiar cómo a partir de la información recopilada en un data center los gobiernos, las agencias de inteligencia y las empresas privadas son capaces de controlar y vigilar a la ciudadanía.

En este apartado me gustaría recalcar que, personalmente, considero que hablar de un data center es hablar de un tipo de archivo. Para justificar mi punto de vista, pasemos a ver primero las definiciones de ambos conceptos.

Según la definición del concepto archivo dada por Ramon Alberch i Fugueras, entendemos por “archivo” un conjunto de documentos recibidos y producidos por personas físicas y jurídicas, públicas o privadas, como resultado de sus actividades, organizados y conservados para poder utilizarlos en la gestión administrativa, la información, la investigación y la cultura. Así mismo, “archivo” designa también a las instituciones responsables que gestionan esta documentación y al espacio físico donde se conserva de manera adecuada para garantizar la accesibilidad y el uso por parte de la ciudadanía.

Por su parte, un data center, o un centro de datos, es una instalación física que aloja servidores y sistemas de almacenamiento donde se ejecutan aplicaciones para procesar, almacenar y analizar grandes cantidades de datos. Los data centers son utilizados por múltiples empresas e instituciones para almacenar y proteger grandes cantidades de información digital corporativa y datos de clientes. Permiten que las organizaciones puedan estar conectadas en todo momento a los datos que allí se gestionan y tener acceso a éstos cuando sea que se requiera.

Ambas definiciones parecen muy diferentes, pero no lo son. Por un lado, en un data center encontramos datos, en lugar de documentos, pero al igual que en un archivo, los datos de un data center son recibidos y producidos por personas físicas y jurídicas, públicas o privadas, y son fruto de sus actividades. El uso de los datos de un data center dependerá del tipo de institución, de manera que es perfectamente posible que se utilicen para fines culturales, para la gestión administrativa y para la investigación.

Por otro lado, un data center es a su vez la institución que gestiona todos estos datos y un espacio físico donde éstos se conservan para garantizar su accesibilidad y uso por parte de usuarios determinados.

Los fines y el uso de los datos de un data center dependerán del tipo de institución de la que estemos hablando, al igual que sucede con un archivo, donde son diferentes en función de si, por ejemplo, es un archivo público de la administración o es el archivo privado de una empresa. En el caso de los data centers, lo más común es que la institución sea una empresa privada, aunque también podemos encontrar data centers en instituciones públicas, como ayuntamientos o agencias de inteligencia.

Carlos Joa, asesor tecnológico de CANTV, la principal empresa de telecomunicaciones venezolana, considera que las características[17] más básicas que debe tener todo data center sin excepción son las siguientes:

- UPS (respaldo de energía interrumpible): sistema que evita la pérdida de información en caso de que se produzca un corte en el suministro eléctrico a través de baterías recargables.
- PDU (unidad de distribución de energía): sistema que distribuye la energía eléctrica a varias computadores a la vez.
- Aire acondicionado, climatización y sistemas de refrigeración adecuados para el correcto mantenimiento y funcionamiento de los equipos informáticos.
- Medidas y sistemas contra incendios adecuadas al material eléctrico
- Sistema CCTV: sistema de video vigilancia cerrado que supervisa múltiples ambientes y actividades a la vez.
- Piso elevado: piso de placas modulares en la parte superior del edificio cuya función es distribuir instalaciones eléctricas y aire acondicionado.
- Sistemas de seguridad y de control de acceso para controlar el acceso a áreas restringidas y evitar la entrada de personal no autorizado al interior del data center. Por ejemplo: tarjetas electrónicas, cerraduras electromagnéticas, detectores de movimiento, bandas magnéticas, escáneres biométricos…
- Dispositivos de control ambiental para detectar cambios repentinos en el ambiente que puedan afectar la integridad de los sistemas.

[17] Joa, Carlos. *Consideraciones para el diseño y la construcción de un data center* [en línea] Slideshare, 2015 [Consulta: 22 febrero 2016]. Disponible en: <http://es.slideshare.net/kacjoa/diseo-y-normas-para-data-centers>.

- Sistemas de protección y distribución de cables de fibra óptica.
- Cableado estructurado según la normativa internacional TI A-942, que establece cinco áreas de distribución diferentes.

Generalmente, todos los grandes servidores de los data centers se suelen concentrar en una sala denominada "sala fría", "nevera" o "pecera". Esta sala requiere un sistema específico de refrigeración para mantener la temperatura entre 21 y 23 grados Celsius, la necesaria para evitar averías y sobrecalentamiento.

Tabla 9. Sala fría del data center de Google en Council Bluffs, Iowa. (Google, 2012)

Algunos de los data centers más conocidos y potentes del mundo son los que tiene Google por todo el globo[18].

Google cuenta con 15 data centers repartidos en diversas localidades de América, Europa y Asia con el objetivo de mantener el funcionamiento de sus productos en perfectas condiciones las 24 horas del día durante los siete días de la semana.

Tabla 10. Edificio del data center de Google de Hamina, Finlandia. (Google 2012)

[18] Google. *Ubicaciones de los centros de datos* [en línea]. 2012 [Consulta: 8 agosto 2016]. Disponible en: <https://www.google.com/about/datacenters/inside/locations/index.html>.

3 Los data centers de Silicon Valley

Los data centers de Silicon Valley engloban a todos los sistemas informáticos y bases de datos masivas de las principales compañías de Internet y telecomunicaciones que recopilan, almacenan, gestionan y tratan datos personales de todos sus usuarios a través de sistemas Big Data.

La importancia de este conjunto de data centers recae en el hecho de que son la fuente principal a la hora de nutrir las grandes bases de datos de algunas de las agencias de inteligencia norteamericanas y europeas. Gracias a todos los datos que los grandes gigantes de las telecomunicaciones y de Internet recopilan y tratan de sus miles de millones de usuarios para sus propios fines e intereses comerciales, los gobiernos y sus agencias de inteligencia nutren a su vez sus propios data centers para intentar ejercer el control total sobre los ciudadanos de sus países a través de la vigilancia masiva.

3.1 ¿De qué data centers hablamos?

En primer lugar, para poder establecer un punto de partida es necesario identificar los principales data centers que hacen uso de tecnologías Big Data y que son el eje principal de la nutrición de las bases de datos de los gobiernos y agencias de inteligencia norteamericanas y europeas.

Evidentemente sería inabarcable tener en cuenta a todas las compañías privadas que hacen tratamiento de datos masivo mediante sistemas Big Data. Sin embargo, solamente unas pocas controlan prácticamente la totalidad de las comunicaciones e Internet a nivel mundial, y gracias a una diapositiva de la Agencia de Seguridad Nacional de los Estados Unidos (NSA) filtrada en 2013 por Edward Snowden, a la que tuvieron acceso los periódicos The Guardian[19] y The Washington Post[20], podemos ver que este reducido número de compañías

[19] Greenwald, Glenn; MacAskill, Ewen. *NSA Prism program taps in to user data of Appe, Google and others* [en línea]. The Guardian, junio 2013 [Consulta: 5 febrero 2016]. Disponible en: <http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data>.

[20] Gellman, Barton; Poitras, Laura. *US intelligence mining data from nine US internet companyies in broad secret program* [en línea]. The Washington Post, junio 2013 [Consulta: 5 febrero 2016]. Disponible en: <https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html>.

todopoderosas, como Google, Facebook y Apple, son las principales proveedoras de datos para las bases de datos de varias agencias gubernamentales.

La diapositiva a la que se hace referencia forma parte de una presentación Power Point de 41 diapositivas clasificada como Top Secret que se utilizó para entrenar a operativos de inteligencia relacionados con el programa PRISM, un programa de la NSA operativo desde 2007 que permite la vigilancia masiva de ciudadanos de todo el mundo mediante un acceso directo a los servidores centrales de empresas estadounidenses líderes en Internet, como Google, Microsoft, Facebook, Yahoo o Apple.

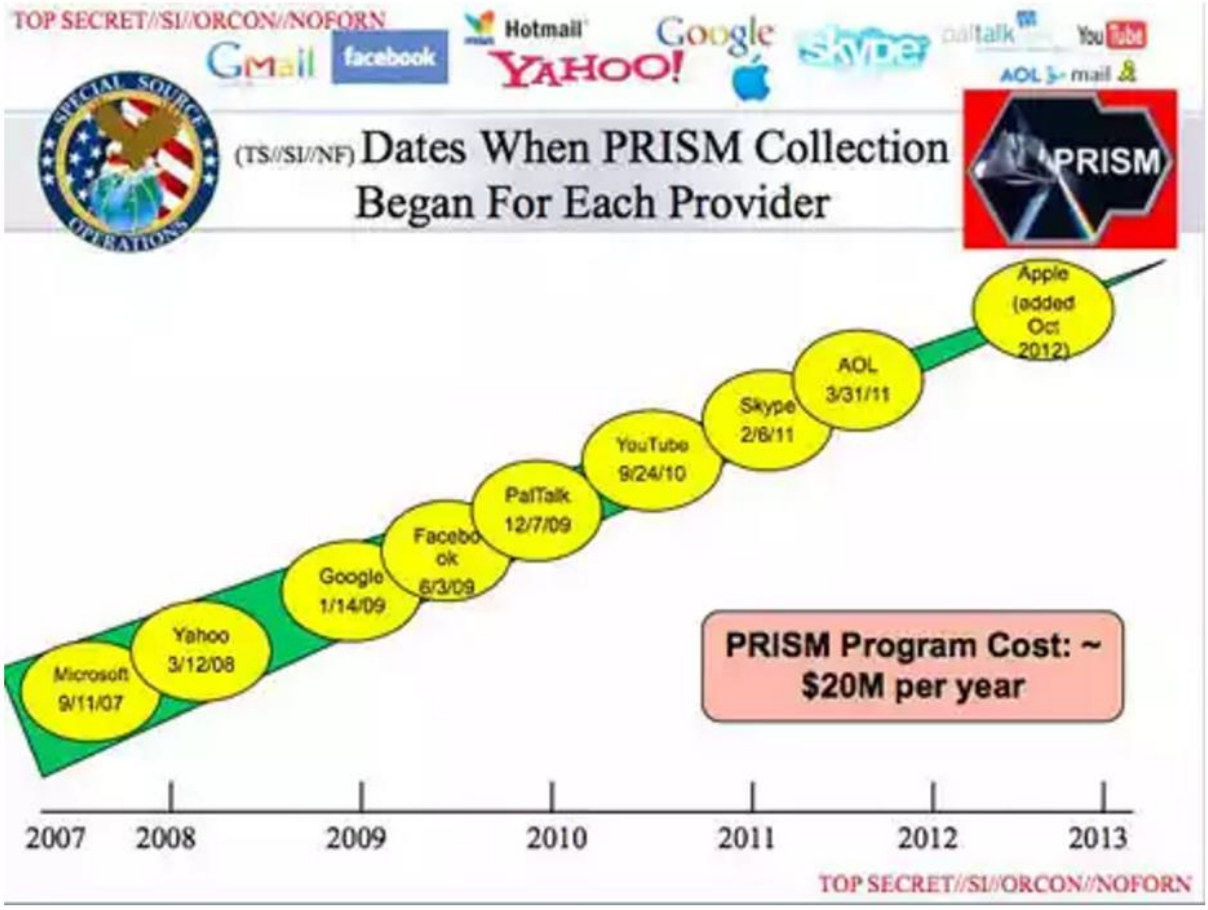

Tabla 11. Proveedores de datos del programa PRISM. (The Guardian, 2013)

Podemos ver que la diapositiva enumera a los proveedores de PRISM hasta el año 2012, algo lógico, puesto que se filtró en el año 2013.

Por otro lado, hay diversos factores a tener en cuenta antes de establecer qué archivos de las grandes compañías señaladas en la lista serán los analizados en este trabajo.

En primer lugar, algunas compañías que aparecen forman parte de otras que también están presentes en la lista. Es el caso de Skype y Youtube, que son propiedad de Microsoft y Google respectivamente. El 10 de mayo de 2011 Microsoft anunció la compra completa de Skype por 8500 millones de dólares[21] y Youtube fue adquirido en octubre de 2006 por Google Inc. a cambio de 1650 millones de dólares.

Si bien es cierto que desde junio de 2015 AOL, empresa que ofrece servicios por Internet, pertenece a la empresa de telecomunicaciones Verizon, AOL aparece en la lista de la diapositiva como proveedor en 2011, antes de pertenecer a Verizon. Por este motivo ambas compañías se estudiaran aparte, ya que también hay pruebas de que Verizon, una de las empresas de telecomunicaciones más importantes de Estados Unidos, facilitara a distintas agencias gubernamentales del país los registros y datos de las llamadas de todos sus clientes.

En segundo lugar, parece ser que la única compañía presente en la lista cuyos datos interceptados realmente podrían ser útiles a la hora de combatir el terrorismo y no ejercer gratuitamente el espionaje indiscriminado es PalTalk. Se trata de una compañía que ofrece a sus usuarios la posibilidad de comunicarse a través de video-llamadas, mensajería instantánea o servicios móviles en la red. Se creó en 1998 y desde entonces atesora unos 4 millones de usuarios de todo el mundo, una cifra ridícula si la comparamos con el billón de usuarios que tiene Facebook.

Según el artículo de la CNN "*Google, Facebook… PalTalk?!*"[22], publicado en junio de 2013, la clave de que PalTalk aparezca en la lista de proveedores de PRISM reside en el hecho de que un considerable número de las salas de chat del servicio son usadas por terroristas y cibercriminales, tal y como afirma la empresa consultora Flashpoint Global Services.

---

[21] El País. *Microsoft compra Skype por 5.920 millones de euros* [en línea]. Barcelona: Ediciones El País S.L., 2011 [Consulta: 24 febrero 2016]. Disponible en: <http://tecnologia.elpais.com/tecnologia/2011/05/10/actualidad/1305018061_850215.html>.

[22] Lobosco, Katie. *Google, Facebook… PalTalk?!* [en línea]. Nueva York: CNN, 2013 [Consulta: 24 febrero 2016]. Disponible en: <http://money.cnn.com/2013/06/07/technology/security/paltalk-nsa-surveillance/>.

Además, según el Washington Post[23], Paltalk tuvo un tráfico considerable durante la Primavera Árabe y la guerra civil en Siria, e indica que según el informe anti-terrorista de las Naciones Unidas de 2009 es también una plataforma con presencia de usuarios relacionados con Al-Qaeda.

Considero pues que en este caso la función de vigilancia para evitar acciones terroristas está muy por encima de la vigilancia masiva indiscriminada a la ciudadanía, por lo que PalTalk no formará parte de los data centers seleccionados.

En tercer lugar, me llama la atención que tres grandes compañías de Internet como son Amazon, Twitter o Dropbox no formen parte de esta lista. Cuando se filtraron las diapositivas del programa PRISM a mediados de 2013, estas ausencias también sorprendieron a la prensa y a gran parte de la opinión pública. Sin embargo, no he encontrado ninguna prueba documental que avale la participación de estas tres compañías como proveedoras de datos del programa de vigilancia PRISM ni de ninguno otro.

De hecho, según un artículo publicado en The Verge en junio de 2013[24], Twitter tiene una relación con el gobierno estadounidense muy poco cooperativa desde su creación, denegando gran parte de sus peticiones de información de datos de usuarios y llevando ante las Cortes aquellas peticiones con las que, aunque la ley las ampare, no están de acuerdo.

De Amazon, más allá de una publicación en el blog de su líder desmarcándose de todo lo que tenga que ver con el programa PRISM y la NSA, no hay ninguna referencia, y en cuanto a Dropbox, la empresa de almacenaje en la nube y servicios de sincronización, la única alusión al tema que nos atañe es que está descrita en las diapositivas del programa PRISM como “coming soon”.

Así pues, tras tomar la lista de las diapositivas de la NSA y tener en cuenta varios factores relacionados, los data centers de las grandes empresas que se analizarán en este trabajo serán los siguientes: Microsoft, Yahoo, Google, Facebook, AOL, Apple y Verizon.

---

[23] Tsukayama, Hayley. *PalTalk: The Prism company that you’ve never heard of* [en línea]. The Washington Post, 2013 [Consulta: 24 febrero 2016]. Disponible en: <https://www.washingtonpost.com/business/technology/paltalk-the-prism-company-that-youve-never-heard-of/2013/06/07/02a0f2c4-cf79-11e2-8f6b-67f40e176f03_story.html>.

[24] Jeffries, Adrianne. *Escape from Prism: how Twitter defies government data-sharing* [en línea]. The Verge, 2013 [Consulta: 2 marzo 2016]. Disponible en: <http://www.theverge.com/2013/6/13/4426420/twitter-prism-alex-macgillivray-NSA-government>.

3.2 Entrada y recopilación de datos

Los datos son la materia prima de cualquier práctica de análisis de negocio. Hasta no hace mucho eso implicaba datos estructurados creados y almacenados por las propias organizaciones. Sin embargo, actualmente, el volumen y los nuevos tipos de datos disponibles en las empresas – y la necesidad de analizarlos en tiempo real para obtener de ellos el máximo valor de negocio – crece vertiginosamente a medida que la casi totalidad de la población mundial accede a las tecnologías de la información.

La comunicación, la rapidez y la interacción social son las principales motivaciones para utilizar estas herramientas. Gracias a una rápida comunicación sin las barreras de tiempo o de distancia, las tecnologías de la información y de telecomunicaciones son el canal más atractivo a la hora de comunicarse e interactuar con otras personas.

La recopilación de la ingente cantidad de datos que tratan y almacenan las grandes compañías mencionadas anteriormente no sería posible sin el uso de tecnologías Big Data. Provienen de fuentes y sistemas muy diversos y diferentes entre sí y prácticamente la totalidad de ellos son datos desestructurados en constante evolución y crecimiento. Todas las compañías que aquí tratamos tienen múltiples programas, software, plataformas, etc. que recopilan diferentes tipos de datos de los usuarios en función del servicio ofrecido y sus características.

Las redes sociales, los buscadores web, la mensajería instantánea o el correo electrónico se han generalizado y la adopción por parte de los usuarios se ha acelerado enormemente en muy poco tiempo. Son canales que reciben una cantidad ingente de todo tipo de datos que ofrecen grandes oportunidades de negocio, tanto en términos de gestión de marca como de apertura a nuevos canales de mercado.

A través de los servicios que ofrecen los gigantes de Internet, los clientes o clientes potenciales se comunican con miles de millones de mensajes, búsquedas e interacciones, informando sobre quiénes son, qué hacen, qué les gusta, qué no, cuáles son sus preferencias en equis tema, qué intereses tienen, qué les preocupa, etc. Gracias a este incesante flujo de datos hay una oportunidad inigualable para que las empresas accedan al cliente y obtengan información personal muy detallada sobre ellos para lograr ventajas competitivas y nuevas oportunidades de negocio.

Pero, ¿cómo extraer valor de semejante cantidad de datos desestructurados? Gracias a las herramientas Big Data estos enormes volúmenes de datos se pueden convertir en un gran recurso a través de la integración de datos. Los datos de los usuarios se pueden analizar y tratar para obtener patrones de comportamiento y establecer segmentos personalizados y tendencias generales.

3.2.1 Datos susceptibles de interés

El principal interés de toda empresa privada es conseguir posicionar su marca e incrementar su visibilidad para obtener más ventas y beneficios, y por supuesto, las compañías que aquí se estudian no son la excepción.

Para poder plantear la mejor estrategia de márquetin posible necesitan fijarse en un determinado tipo de datos para poder tratarlos y extraer valor de ellos. Pero, ¿cuáles son estos datos? Los datos que más interés despiertan a este tipo de empresas son múltiples: *likes* en las redes sociales, descargas, visionado de contenido, visionado de anuncios, links abiertos, palabras clave utilizadas, momento y lugar en el que se usan esas palabras clave, tráfico, transacciones, tipos de consultas, datos de geolocalización...

Por ejemplo, si analizamos el ciclo que siguen las empresas y organizaciones a la hora de establecer sus campañas comerciales en Facebook, se puede ver que, en primer lugar, les interesan todos aquellos datos que estén relacionados con las actividades que realizan los usuarios dentro de la red social: qué utilizan, con qué frecuencia, por qué motivos...

El sistema apenas varía si hablamos de buscadores o plataformas digitales de cualquier tipo, de manera que podríamos afirmar que lo primero que priorizan las grandes compañías de Internet y telecomunicaciones es focalizar su actividad en los grupos potencialmente más importantes para ellas, es decir, determinar grupos de interés. Es lo que se conoce como segmentación de usuarios.

Una vez determinados los distintos grupos de interés en los cuales se focalizaran diferentes campañas comerciales, será necesario recabar datos acerca de sus intereses, gustos, preferencias y costumbres para poder decidir qué contenidos difundir en cada caso y en qué tono. Es decir, establecer anuncios personalizados y únicos para cada usuario.

Para ello, deberán ser datos que respondan a las siguientes preguntas: ¿de qué habla el grupo de usuarios de interés?, ¿qué les interesa?, ¿qué contenidos están más predispuestos a compartir con otros usuarios? Para que la estrategia sea lo más eficaz posible, las herramientas Big Data extraerán y analizarán estos datos a partir del rastro digital que dejen los usuarios en Internet.

Un ejemplo de los datos susceptibles de interés lo podemos ver claramente en la plantilla que Facebook ofrece a sus anunciantes para facilitarles la segmentación de los usuarios (que realiza el propio Facebook), donde hay datos demográficos, datos sobre intereses y preferencias, temas, categorías, palabras clave o personas a las que les gusta una página, incluidos temas por productos, marcas, religión, salud o ideología política.

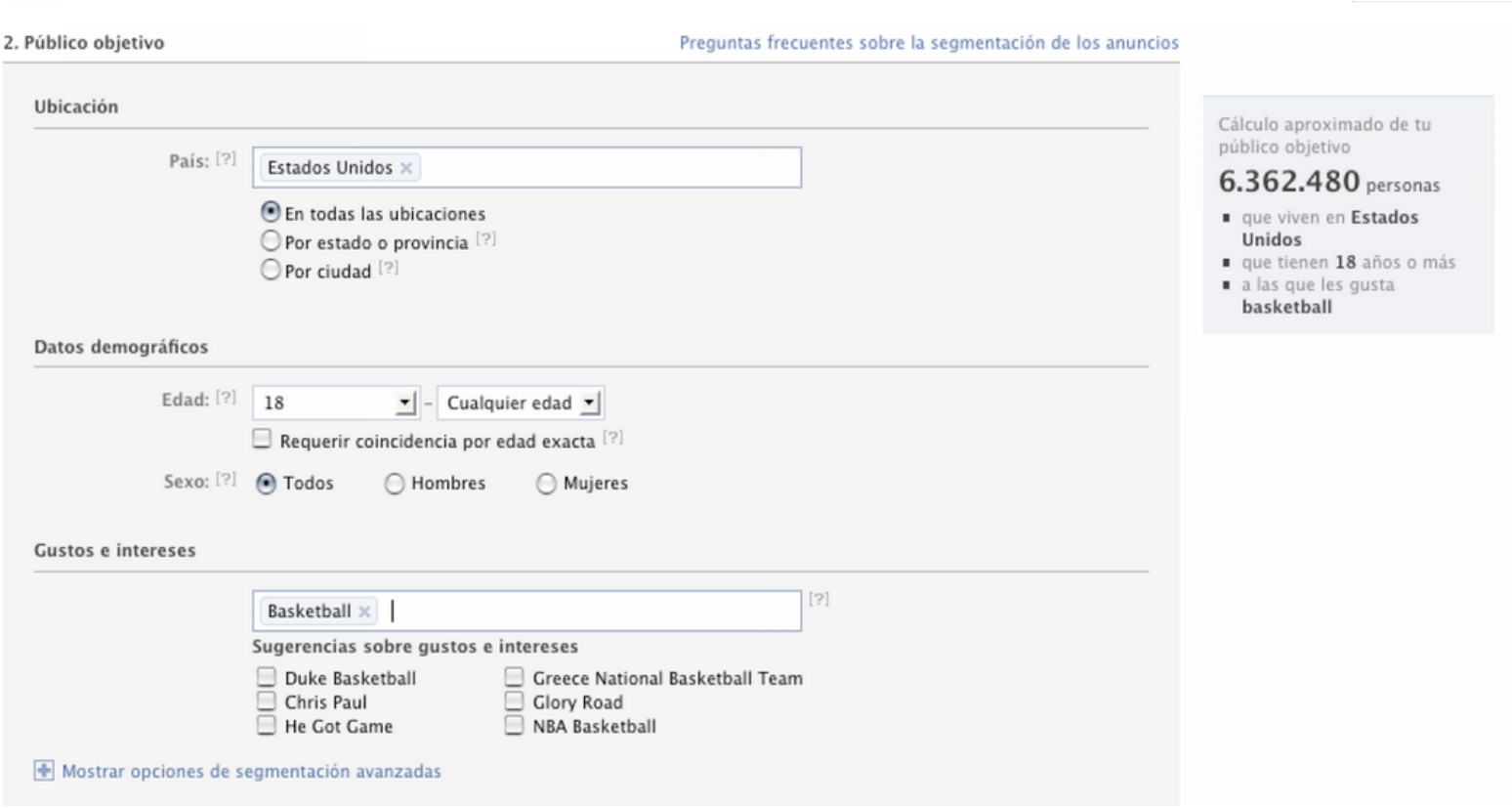

Tabla 12. Preguntas frecuentes sobre la segmentación de los anuncios. (Facebook, 2014)

3.2.2 Entrada y recopilación de datos

Los datos que capturan los grandes archivos del Big Data provienen de nosotros mismos. Los fabricamos directa e indirectamente segundo tras segundo a través de los productos que nos ofrecen. Para hacernos una idea del gran volumen de datos que puede recopilar uno solo de esos productos, un IPhone en la actualidad tiene mucha más capacidad de cómputo que la NASA cuando llegó el hombre a la Luna.

Según el informe del Grupo TRC "*Conceptos básicos de Big Data*"[25], la procedencia de los datos se puede clasificar según las siguientes categorías:

Generados por las personas: El hecho de enviar correos electrónicos por e-mail o mensajes por WhatsApp, publicar un estado en Facebook, tuitear contenidos o consultar algo en Google son cosas que hacemos a diario y que crean nuevos datos y metadatos que pueden ser analizados. Se estima que cada minuto al día se envían más de 200 millones de e-mails, se comparten más de 700.000 piezas de contenido en Facebook, se realizan dos millones de búsquedas en Google o se editan 48 horas de vídeo en YouTube.

Transacciones de datos: La facturación, las llamadas o las transacciones entre cuentas generan información que, tratada de cierta manera, puede convertirse en datos relevantes. En las transacciones bancarias, lo que para el usuario es un simple ingreso de equis euros, la computación lo interpretará como una acción llevada a cabo en una fecha y un momento determinado, en un lugar concreto, entre unos ciertos usuarios registrados, etc.

E-marketing y web: Cuando navegamos por internet es cuando generamos más cantidad de datos, ya que desde hace varios años los mismos usuarios son creadores de contenido gracias a la posibilidad de interacción con cualquier sitio web. Existen muchas herramientas de tracking utilizadas en su mayoría para el marketing y el análisis de negocio, como los movimientos de ratón, que se quedan grabados en mapas de calor, o los registros de cuánto tiempo pasamos en cada página web y cuándo y por qué las visitamos.

[25] TRC. *Conceptos básicos de Big Data* [en línea]. Madrid: TRC, 2014 [Consulta: 28 marzo 2016] Disponible en: <http://www.trc.es/pdf/descargas/big_data.pdf>.

Machine to Machine (M2M): Son las tecnologías que comparten datos con otros dispositivos, como los medidores y sensores de temperatura, luz, presión, sonido… Transforman las magnitudes físicas o químicas y las convierten en datos. Existen desde hace décadas, pero la llegada de las comunicaciones inalámbricas (Wi-Fi, Bluetooth, RFID…) ha revolucionado el mundo de los sensores, y con la gran inversión que se está realizando en transformar las ciudades convencionales en Smart Cities, dentro de algunos años cobrarán mucho más protagonismo como fuentes de datos masivos útiles.

Biométrica: Son el conjunto de datos que provienen de la seguridad, defensa y servicios de inteligencia, generados por lectores biométricos como escáneres de retina, escáneres de huellas digitales o lectores de cadenas de ADN. El propósito de estos datos es proporcionar mecanismos de seguridad y suelen estar custodiados por organismos de defensa y departamentos de inteligencia.

Los grandes data centers que se analizan en el presente trabajo extraen los datos se sus usuarios a partir de todas estas categorizaciones, a excepción de la biométrica y en menor medida de la M2M. La captura de datos de las grandes compañías de Internet se basa principalmente en el contenido originado por las propias personas, en las transacciones y en la web, y las grandes compañías de telecomunicaciones en las transacciones.

El caso es, ¿qué ocurre cuándo nuestros gigantes de Internet encuentran las fuentes de los datos que pretenden? Que muy posiblemente tengan tablas de origen sin estar relacionadas, por lo que el principal objetivo de la captura y recopilación de datos es recoger los datos en un mismo lugar y darles un formato.

Aquí es donde entran en juego las plataformas ETL (Extract, Transform and Load). Básicamente, una plataforma ETL se encarga de extraer datos de diferentes fuentes y sistemas, transformarlos en datos útiles y manejables y cargarlos en la base de datos correspondiente, que posteriormente será la encargada de analizarlos, gestionarlos y recuperarlos.

Así pues, una plataforma ETL opera en tres fases: extracción, transformación y carga. En la extracción se extraen datos de calidad, tanto de fuentes homogéneas como heterogéneas, en la transformación se aplican una serie de reglas para transformar los datos de acuerdo a unos requerimientos (por ejemplo, que las medidas tengan la misma dimensión o que se apliquen reglas de validación avanzadas) y la carga el ETL se asegura de cargar los datos de la forma más segura y eficiente posible.

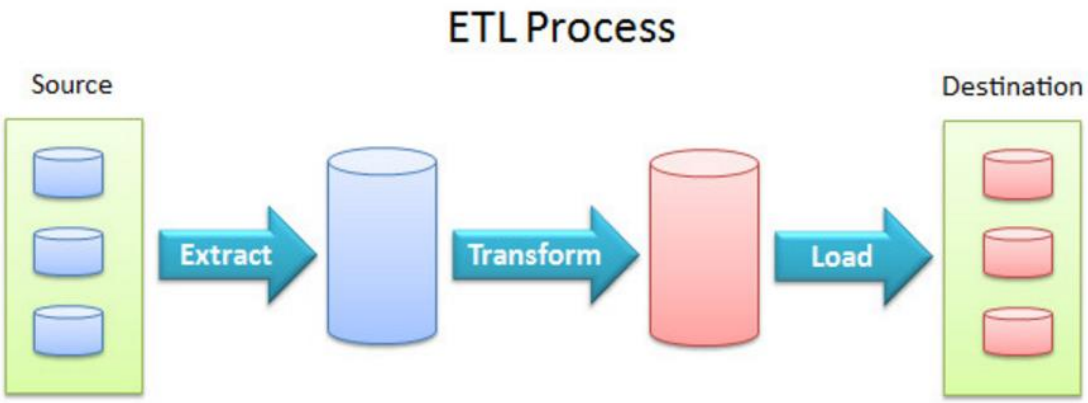

Tabla 13. Proceso de un ETL. (Channelbiz, 2015)

Según los datos de la encuesta realizada por Xplenty[26] en 2015, un proveedor de plataformas de integración, el coste de los procesos ETL cuando se trabaja con Big Data es muy elevado, hasta el punto de que un tercio de los profesionales gasta entre un 50% y un 90% de su tiempo en limpiar los datos que luego van a analizar. Sin embargo, no creo que este sea un hecho que preocupe a los grandes gigantes de Internet y de telecomunicaciones que se analizan en este trabajo.

3.2.3 Almacenamiento

Tal y como se explica en el apartado de conceptos básicos, el almacenamiento de los datos mediante tecnologías Big Data se realiza a través de bases de datos NoSQL, que procesan grandes cantidades de datos semiestructurados y desestructurados. Existen diferentes grupos de bases de datos NoSQL, pero las que nos atañen son las siguientes: key-value, documentales, orientadas a columnas y de grafos.

[26] Full session – Big Data's "janitor" problem – Is it killing ROI? 14 de mayo de 2015. Acceso el 2 de abril de 2016. Vídeo de Youtube disponible en: <https://www.youtube.com/watch?v=YL5TK0vbuhc&feature=youtu.be>.

Base de datos key-value

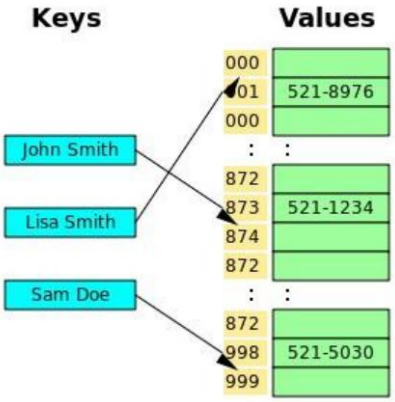

Tabla 14. Base de datos Key-Value. (Telefónica, 2014)

Las bases de datos Key-value son el modelo NoSQL más popular, además de ser la más sencilla en cuanto a funcionalidad. En este tipo de sistema, cada elemento está identificado por una llave única, lo que permite la recuperación de la información de forma muy rápida. No tiene relaciones ni estructura, cosa que hace que la gestión sea muy rápida y eficiente en sistemas distribuidos.

El gran inconveniente de este tipo de base de datos NoSQL es que muchos nodos, es decir, grupos de datos y objetos, no pueden ser fácilmente modelados en el sistema clave-valor por la falta de relaciones.

El ejemplo más conocido de Key-value es Apache Cassandra, utilizado por Apple y Facebook. Tal y como se reveló en el Cassandra Summit San Francisco de 2014, Apple utiliza alrededor de 75,000 nodos de Cassandra. En el 2008, Facebook utilizaba más de 200 nodos solamente para su sistema de búsquedas en la bandeja de entrada[27]. Imaginemos cómo se habrá multiplicado esa cifra en la actualidad.

[27] Lakshman, Avinash. *Cassandra: a structured storage system on a P2P Network* [en línea]. Facebook Engineering, 2008 [consulta: 23 febrero 2016]. Disponible en: <https://www.facebook.com/note.php?note_id=24413138919&id=9445547199&index=9>.

Base de datos documental

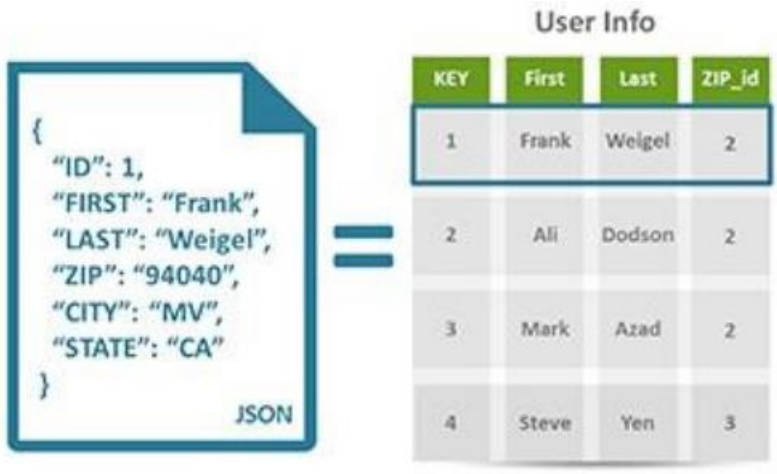

Tabla 15. Base de datos documental. (Telefónica, 2014)

Se trata de una base de datos basada en la Key-value, con la diferencia de que en este caso los datos se almacenan en un formato que la base de datos puede entender o aceptar. Los datos se representan en colecciones. Una colección define a un grupo de documentos similares en esquema libre, es decir, documentos de distintos tipos. Por lo tanto, dentro de una colección se pueden almacenar documentos de tipologías y estructuras diferentes.

La gran ventaja de este tipo de base de datos es que cuentan con todos los beneficios de las key-value sin la limitación de consultar solamente por clave, ya que los datos se almacenan en un formato comprensible para la máquina, cosa que permite consultas avanzadas sobre el contenido de un documento.

Son las bases de datos NoSQL más versátiles y se pueden utilizar en gran cantidad de proyectos, incluyendo muchos que tradicionalmente funcionarían sobre bases de datos relacionales.

Los ejemplos más destacados son algunas de las bases de datos desarrolladas mediante la tecnología Hadoop, de Apache. Hadoop es una familia de tecnologías de código abierto supervisadas por Apache Software Foundation, y por ello algunos de sus productos permiten varias combinaciones y se encuentran en paquetes comercializados.

Tal y como podemos ver en la wiki de Hadoop[28], AOL, Facebook, Yahoo y Microsoft son fieles usuarios de bases de datos documentales basadas en Hadoop.

AOL utiliza Hadoop no solamente como base de almacenamiento de datos NoSQL, sino también como plataforma ETL. Combina la plataforma ETL basada en Hadoop con algoritmos avanzados, y de esta manera la compañía puede extraer datos de una manera que permite desde el principio la segmentación de usuarios y su análisis de comportamiento.

Por su parte, Facebook utiliza bases de datos documentales de Hadoop para almacenar copias de registros internos y grandes fuentes de datos, y Yahoo! para almacenar sus sistemas de anunciantes y búsquedas en la web.

Por otro lado tenemos a Microsoft, que está detrás de muchos de los proyectos de Hadoop, tal y como se refleja en su página sobre Microsoft Azure[29]. Microsoft Azure es un sistema de almacenaje y análisis de datos en la nube basado al 100% en los sistemas Hadoop, por lo que lo más probable es que se trate de una base de datos NoSQL híbrida, no solamente documental.

Base de datos orientada a columnas

Son muy parecidas a las bases de datos documentales, ya que en ambos tipos los datos se almacenan como una columna de valores por fila con su campo clave. Es decir, se almacenan columnas enteras en una sola fila. La principal diferencia es que, mientras que la documental permite almacenar información compleja de forma flexible, la orientada a columnas tiene un formato más fijo, basado en columnas y sub-columnas.

Son bases de datos pensadas y diseñadas para almacenar ingentes cantidades de datos, ya que es cuando más aumenta su rendimiento. De hecho, es un sistema que se usa en casos donde las claves contengan como mínimo más 100 atributos. Además, el servidor es capaz de realizar muchas operaciones al poder acotar los datos de una fila entera, ordenar las columnas automáticamente o hacer mapeos a listas de datos.

[28] Hadoop Wiki. *Powered by Apache Hadoop* [en línea]. Última actualización 8 diciembre 2015 [Consulta: 23 febrero 2016]. Disponible en: <http://wiki.apache.org/hadoop/PoweredBy>.
[29] Microsoft Azure. *Hadoop. ¿Qué es Hadoop?* [en línea]. Microsoft Corporation [Consulta: 23 febrero 2016]. Disponible en: <https://azure.microsoft.com/es-es/solutions/hadoop/>.

La masiva base de datos de Google, la BigTable[30], es la precursora por excelencia de las bases de datos orientadas a columnas. Es como un mapa multidimensional con tres dimensiones: filas, columnas y marca temporal. Los datos se dividen en columnas que contienen tablas compuestas por celdas. Cada celda tiene una marca temporal que permite visualizar la evolución del dato que contenga a lo largo del tiempo.

Utiliza un sistema de archivos del propio Google, el Google File System (GFS), que en el 2015 realizaba la compresión y la descompresión de datos a través de dos algoritmos muy veloces: 100-200 MB/s para comprimir y 400-1000 MB para descomprimir. Es muy probable que desde el pasado año su velocidad haya aumentado.

Además de la BigTable de Google, destaca la HBase de Hadoop, una potente base de datos interactiva orientada en columnas especialmente diseñada para poder alcanzar varios Petabytes de capacidad y manejar millones de solicitudes por segundo. Facebook la utiliza para gestionar las relaciones existentes entre los diferentes perfiles de usuarios.

Base de datos de grafos

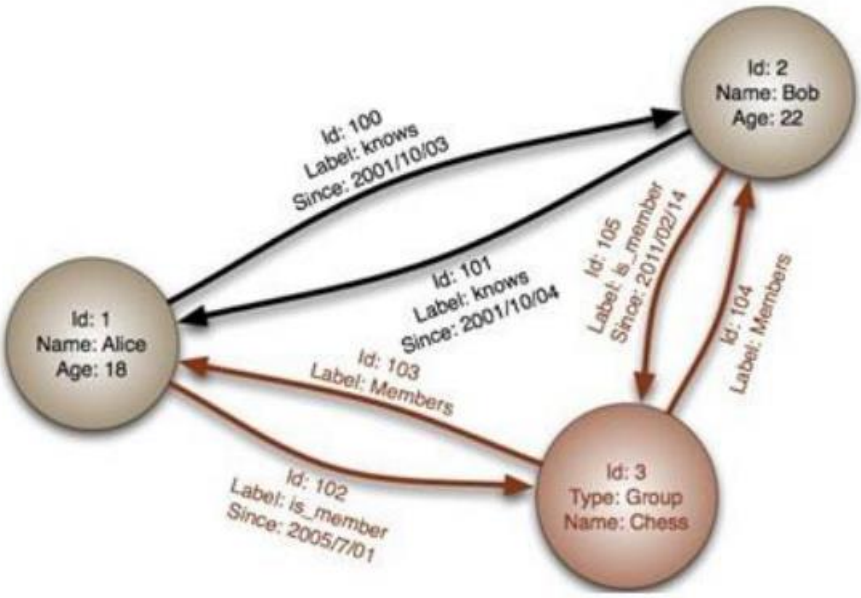

Tabla 16. Base de datos de grafos. (Telefónica, 2014)

[30] BBVA Open4U. *Bigtable, el servicio de base de datos NoSQL con el que Google quiere dominar los Big Data* [en línea]. BBVA, 2015 [Consulta: 8 abril 2016]. Disponible en: <http://www.bbvaopen4u.com/es/actualidad/bigtable-el-servicio-de-base-de-datos-nosql-con-el-que-google-quiere-dominar-los-big-data>.

Estas bases de datos, más allá del hecho de ser de tipo NoSQL, no tienen nada que ver con las anteriores, ya que rompen con la idea de tablas y columnas y se basan en la teoría de grafos. La teoría de grafos establece que la información está localizada en nodos, y que las relaciones entre los nodos son aristas. Los nodos son como entidades y las aristas relaciones entre entidades.

Las bases de datos orientadas a grafos ayudan a encontrar relaciones entre grandes cantidades de datos y dar sentido al puzzle completo. La mayoría son de esquema libre, es decir, sin estructura definida, y almacenan el contexto de los nodos y las aristas a través de listas adyacentes. Con una lista de adyacentes, el nodo indica en qué otros nodos tiene aristas.

Son las bases de datos NoSQL menos utilizadas, aunque se usan mucho en servicios que requieren datos de geolocalización o para encontrar amigos en común en redes sociales. Una de las más conocidas es Neo4j, y aunque no haya constancia de que ninguno de los grandes gigantes de Internet que se estudian en este trabajo la utilice, podemos ver a empresas muy importantes que sí lo hacen, como eBay, Hewlett-Packard o Lufthansa.

3.3 Análisis de los datos

Después de haber visto cómo los gigantes de Internet recopilan y almacenan grandes conjuntos de datos generados por sus usuarios en sus bases de datos NoSQL, es necesario saber cómo los analizan.

Aunque actualmente se están desarrollando nuevas técnicas para mejorar las establecidas y es un campo en constante evolución, de momento las más destacables son las siguientes: pruebas A/B, data mining, inteligencia artificial, procesamiento de señales, fusión e integración de datos, optimización, análisis de redes, análisis espacial y simulación.

Pruebas A/B

Las pruebas A/B consisten en comparar a un grupo controlado con otros de prueba para determinar los tratamientos que harán que mejore un objetivo variable dado. A las empresas, las pruebas A/B les permiten probar con visitantes reales diferentes propuestas de diseño para identificar cuál de ellas genera más tasas de conversión o más ventas. Es decir, las pruebas A/B permiten establecer de manera objetiva el mejor diseño para los sitios web de los anunciantes de los grandes gigantes de Internet.

En este caso, sus aplicaciones son diversas. Para decidir qué anuncios se promocionarán en redes sociales, buscadores y páginas web, las pruebas A/B permiten determinar de entre varias páginas de inicio cuál de ellas generará más ventas, qué diseño de botón de acción (color, tamaño, forma…) generará más clics, qué disposición de espacio en el sitio web generará más usuarios o qué tipo de diseño generará más negocio o disminuirá los porcentajes de rebote de los usuarios.

Data mining

El data mining o la minería de datos consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información es previamente desconocida y puede resultar útil para algún proceso, por ejemplo, para incrementar la satisfacción de la experiencia de los usuarios. La minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Para el data mining no son los datos en sí lo más relevante, sino la información que se encierra en sus relaciones, fluctuaciones o dependencias.

Las aplicaciones del data mining en redes sociales y servicios de Internet permiten desde determinar qué segmentos de la población responderán mejor a una oferta hasta establecer modelos de comportamientos de compra de clientes. Es una técnica vital para la publicidad personalizada, ya que además de los datos demográficos más comunes como el sexo, la edad, el empleo o el lugar de residencia, se añaden parámetros como aficiones, webs que se visitan, grupos a los que se pertenecen, personas con las que se habla, intereses, e incluso el análisis semántico de lo que decimos.

Dentro de la minería de datos destacan cuatro técnicas específicas: las reglas de asociación, la clasificación, el análisis de clústeres y la regresión.

Las reglas de asociación son un conjunto de técnicas usadas para descubrir relaciones de interés entre las variables de las grandes bases de datos NoSQL. Se basa en una variedad de algoritmos que generan y testean posibles reglas: si un usuario compra tal hay tantas posibilidades que también compre tal, si un usuario consulta tal hay tantas posibilidades de que también consulte aquello otro, etc.

Las técnicas de clasificación identifican las categorías donde hay nuevos puntos de datos basándose en puntos de datos que anteriormente ya han sido categorizados. Se utiliza principalmente para predecir comportamientos de segmentos de usuarios específicos. En este sentido destaca el reconocimiento de patrones, que extrae la información de objetos tanto físicos como abstractos para establecer una serie de propiedades de los conjuntos de dichos objetos. Esta información extraída es la que se descarta de lo que el sistema considera como información redundante o irrelevante.

Para entender mejor el reconocimiento de patrones sólo hace falta fijarnos en Facebook y en su sistema de reconocimiento facial. Actualmente, Facebook cuenta con Deepface[31], que se encarga de mejorar el sistema de etiquetado de las fotografías. Según el informe de Deepface, el software crea modelos 3D de los rostros que aparecen en fotografías y los analiza a través de un sistema de aprendizaje que imita la estructura de las neuronas en el cerebro para dibujar conexiones. Es decir, lo que hace es realizar un análisis de grafos. Su grado de precisión es del 97,25%, apenas unas décimas por debajo del que tiene el ojo humano.

En cuanto al análisis de clústeres, se trata de un método estadístico de clasificación de objetos. Se basa en dividir grupos muy diversos en grupos más pequeños que contengan objetos similares, cuyas características de similitud no se conocen de antemano. Un ejemplo sería segmentar usuarios en base a grupos que contengan usuarios similares.

[31] Taigman, Yaniv [et.al]. *DeepFace: Closing the Gap to Human-Level Performance in Face Verification* [en línea]. Facebook, 2014 [Consulta: 17 abril 2016]. Disponible en: <https://www.facebook.com/publications/546316888800776/>.

Por su parte, la regresión es un conjunto de técnicas estadísticas que determinan cómo el valor de una variable cambia cuando otras variables independientes también cambian. Destaca especialmente su uso en la predicción de resultados, como por ejemplo las famosas búsquedas sugeridas de Google.

Inteligencia artificial

La inteligencia artificial se basa en una serie de algoritmos capaces de evolucionar por sí mismos basándose en datos empíricos. El aspecto más importante que cubre la inteligencia artificial en el análisis de datos es el aprendizaje automático a la hora de reconocer patrones muy complejos y ser capaz de tomar decisiones inteligentes en consecuencia.

Un ejemplo clarísimo de inteligencia artificial es el procesamiento de lenguaje natural, un conjunto de técnicas computacionales y lingüísticas que, a través de algoritmos, analizan el lenguaje natural. Gracias al procesamiento del lenguaje natural se pueden realizar análisis de sentimientos para determinar reacciones: reacciones de clientes potenciales ante una campaña comercial, reacciones de usuarios ante un cambio en una aplicación o servicio…

Procesamiento de señales

Se entiende por procesamiento de señales al conjunto de técnicas de ingeniería eléctrica y de matemáticas aplicadas que analizan señales continuas y discretas, como señales de radio, sonidos o imágenes.

El procesamiento de señales permite realizar una lectura más precisa mediante la combinación de datos de un conjunto de fuentes de datos menos precisos, es decir, extraer el ruido de la señal. Se ha aplicado para reconocer qué roles asume la gente en la producción de datos del entorno o para registrar las interacciones vocales que se producen entre grandes grupos de individuos con teléfonos inteligentes.

Fusión e integración de datos

Mediante un conjunto de técnicas combinadas, como las dos anteriores, capaces de integrar y analizar datos procedentes de múltiples fuentes, se pueden extraer datos mucho más precisos que si solamente se analizan los de una sola fuente.

Por ejemplo, las técnicas utilizadas en el procesamiento de señales pueden usarse para la fusión de datos a través de sensores combinados en el Internet de las cosas para desarrollar perspectivas más integradas. También los datos procedentes de redes sociales y foros analizados mediante el procesamiento de lenguaje natural pueden combinarse con datos de ventas a tiempo real para poder determinar qué efecto está teniendo una campaña de márquetin en los sentimientos de un usuario y su comportamiento de compra.

Optimización

La optimización se basa en técnicas numéricas para rediseñar sistemas y procesos complejos, para así poder mejorar el rendimiento, los procesos operativos o facilitar la toma de decisiones estratégicas de una organización. Un ejemplo lo tenemos en los denominados algoritmos genéticos, capaces de combinarse entre sí o mutar en función del entorno.

Análisis de redes

Un conjunto de técnicas para caracterizar las relaciones que hay entre los nodos que pasan desapercibidos en una red. Se estudian las asociaciones, las relaciones y los flujos entre las personas, grupos, organizaciones o sitios web. El análisis de redes permite analizar relaciones humanas, como las conexiones entre individuos, o ver quién tiene mayor influencia sobre quién, así como también analizar cómo viaja la información.

Un ejemplo de la aplicación del análisis de redes lo encontramos en la empresa de monitorización y análisis de redes sociales Sentisis, que a través de esta técnica determinó las claves para entender cómo llegarían las aficiones del Atlético de Madrid y del Real Madrid a la final de la Champions League de mayo de 2014, los puntos desde dónde se apoyarían a cada conjunto y el equipo declarado como favorito.

Análisis espacial

Es un conjunto de técnicas, en su mayoría estadísticas, que analizan las propiedades topológicas, geométricas y geográficas codificadas en un conjunto de datos. Estos datos suelen provenir de los sistemas de información geográfica.

Las principales aplicaciones del análisis espacial van desde determinar cuán dispuesto está un consumidor a comprar un producto según su ubicación, hasta establecer la mejor manera de instaurar una cadena de suministros con bases en diferentes lugares. Gracias al análisis espacial, se puede generar márquetin viral, seguir el comportamiento de los usuarios, identificar y obtener información cuantitativa de los usuarios, generar *feedback* constante, posibilitar la medición del tráfico del negocio…

Simulación

La simulación es una técnica que consiste en modelar el comportamiento de sistemas complejos de predicción y planificación, como la analítica predictiva. Gracias a la simulación es posible realizar histogramas que proporcionen una distribución de probabilidad de resultados, algo fundamental para las empresas, ya que la clave de los negocios es saber cómo va a comportarse el cliente frente al lanzamiento de productos y campañas de márquetin. Qué mejor manera de entender la realidad que acercándose a ella y haciendo simulaciones teniendo en cuenta su complejidad.

3.4 Explotación de los datos

O lo que es lo mismo, ¿qué van a hacer los grandes gigantes de Internet y de las telecomunicaciones con todos los grandes conjuntos de datos almacenados y analizados? Como ya se ha mencionado, el objetivo de toda empresa privada es generar beneficios. En este caso, los gigantes que se analizan en este trabajo principalmente viven de un modelo de negocio basado en el análisis de los datos de sus usuarios para ofrecerles la publicidad y los servicios que mejor encajen con ellos.

Si nos centramos en las formas de explotación más comunes podemos determinar las siguientes:

Recomendaciones y consejos: los vendedores online pueden hacer uso de los datos analizados mediante tecnología Big Data para recomendarse entre ellos o para aconsejar a los clientes sobre otros productos y servicios que podrían interesarles a partir del análisis del perfil de usuario y de su comportamiento online. Por ejemplo, en las redes sociales vendría a ser el cásico “gente que podrías conocer”, en Google el “quizás querías decir”, y en la venta de productos el “productos relacionados” o “otros clientes también compraron”.

Determinar sentimientos: las herramientas de análisis de texto avanzadas analizan el texto no estructurado para que las empresas puedan determinar los sentimientos de los usuarios en relación a sus marcas, la propia empresa o los productos y servicios concretos que ofrece.

Analizar campañas de márquetin: los departamentos de márquetin de todos los sectores monitorizan y determinan la efectividad de sus campañas a partir de grandes volúmenes de datos con granularidad incremental, como datos sobre el flujo de los clics de un usuario. De esta manera se consigue aumentar la precisión del análisis.

Adelantarse al abandono de clientes: gracias al análisis de los datos relacionados con el comportamiento de los clientes se pueden identificar patrones que indiquen cuáles de ellos son más susceptibles de abandonar la compañía en detrimento de un producto o servicio de la competencia. De ese modo, pueden avanzarse a los hechos y tomar medidas para evitar el posible abandono de clientes valiosos.

Determinar los clientes más importantes: a raíz del análisis de datos se puede determinar qué usuarios tienen mayor influencia sobre otros. Esto ayuda a las compañías a determinar quiénes son sus “clientes más importantes”, que no son siempre los que más productos consumen o más dinero gastan, sino los que poseen una mayor capacidad de influencia sobre el comportamiento de compra del resto, como los *youtubers* o los blogueros.

Optimizar las experiencias de cliente: mediante el análisis de los datos se puede integrar la información recogida de distintas fuentes para obtener una visión completa de la experiencia de cliente. De este modo, las empresas pueden comprender el impacto que tiene un canal de interacción con los clientes sobre otro de cara a optimizar el ciclo de vida completo de la experiencia de cliente.

4 Los data centers de la vigilancia masiva

"*La industria de la vigilancia trabaja mano a mano con gobiernos de todo el mundo para ayudarles a espiar de forma ilegítima a sus ciudadanos. Con escasa vigilancia y ninguna regulación efectiva, nos vemos involucrados en este tipo de espionaje sin límites en contra de nuestra voluntad y, frecuentemente, sin enterarnos*".

Julian Assange, fundador de Wikileaks.

Actualmente, las principales potencias democráticas del mundo se han dotado de múltiples y efectivos sistemas y mecanismos de control de comunicaciones que han llegado a tener a día de hoy un alcance masivo y planetario, una constante que se buscó con ahínco especialmente desde el 11-S y la promulgación de la Patriot Act en Estados Unidos en pos de la lucha contra el terrorismo.

Empresas filiales de agencias como la CIA o la NSA desarrollaron hace pocos años una nueva generación de herramientas tecnológicas centradas en la creación de programas dedicados al control de las comunicaciones electrónicas y telefónicas y de los miles de millones de datos que generan a partir de técnicas como el data mining o el procesamiento del lenguaje natural. Hablamos de nuevo del Big Data, toda una revolución también en materia de gobierno e inteligencia. Gracias a este fenómeno, los servicios de inteligencia son más omnipresentes que nunca.

Gracias a las filtraciones de Edward Snowden en junio de 2013, hemos podido conocer este tipo de praxis por parte de los principales gobiernos democráticos y sus agencias de inteligencia. La prensa se hizo especial eco de programas de vigilancia masiva como PRISM y el abuso al respecto por parte de agencias norteamericanas como la NSA, pero no son, ni mucho menos, el único mecanismo de espionaje y captura de datos, ni la única agencia implicada en la vigilancia masiva de la ciudadanía global.

4.1 El informe Moraes

Poco después de las primeras revelaciones por parte de Edward Snowden, las Naciones Unidas y la Organización de Estados Americanos manifestaron su gran preocupación al respecto e instaron inmediatamente a las autoridades correspondientes a revisar su legislación y modificar las prácticas intrusivas que se habían filtrado, para asegurar el cumplimiento de los derechos humanos. Ambos organismos lo reflejaron a través de una Declaración conjunta sobre programas de vigilancia y su impacto en la libertad de expresión[32] el 21 de junio de 2013.

Tras la declaración de la ONU y la Organización de los Estados Americanos, el Parlamento Europeo (PE), aprobó el 4 de julio de 2013 una resolución en la que encargaba a su Comisión de Libertades Civiles, Justicia y Asuntos de Interior (LIBE) una investigación exhaustiva de los programas de vigilancia masiva. El 21 de febrero del 2014 el LIBE presentó el *Informe sobre el programa de vigilancia de la Agencia Nacional de Seguridad de los EEUU, los órganos de vigilancia en diversos Estados miembros y su impacto en los derechos fundamentales de los ciudadanos de la UE y en la cooperación transatlántica en materia de Justicia y Asuntos de Interior*, más conocido como Informe Moraes[33].

Lo más significativo del Informe Moraes es que reconoce la existencia de programas secretos de vigilancia que no están relacionados únicamente con cuestiones de seguridad nacional, que atentan contra los derechos fundamentales de todos los ciudadanos de manera indiscriminada, sin basarse en sospechas, y que actúan contra la seguridad y la fiabilidad de las redes de comunicación. Así mismo, reconoce la existencia del uso de tecnologías intrusivas muy avanzadas por parte de los Estados Unidos y de varios Países Miembros capaces de recopilar y analizar los datos de las comunicaciones de todo el mundo.

---

[32] Organización de los Estados Americanos. *Declaración conjunta sobre programas de vigilancia y su impacto en la libertad de expresión* [en línea]. Washington DC: OEA, 2013 [Consulta: 20 abril 2016]. Disponible en: <http://www.oas.org/es/cidh/expresion/showarticle.asp?artID=927&>.

[33] Comisión de Libertades Civiles, Justicia y Asuntos de Interior. *Informe sobre el programa de vigilancia de la Agencia Nacional de Seguridad de los EEUU, los órganos de vigilancia en diversos Estados miembros y su impacto en los derechos fundamentales de los ciudadanos de la UE y en la cooperación transatlántica en materia de Justicia y Asuntos de Interior* [en línea]. Parlamento Europeo, 2013 [Consulta: 24 abril 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A7-2014-0139+0+DOC+XML+V0//ES>.

Así pues, la investigación del LIBE pone de manifiesto que las prácticas secretas de vigilancia masiva a través de tecnología invasiva y de alcance mundial existen, y que además, son llevadas a cabo por los principales países democráticos del mundo, tal y como se puede leer en el punto 22 del apartado de Principales Conclusiones, donde se pide a los Estados miembros que forman parte de los "Nueve Ojos" y "Catorce Ojos" – de los que hablaremos en el siguiente apartado del trabajo – que revisen sus legislaciones y las prácticas de sus agencias de inteligencia para garantizar que se respetan los principios de legalidad, los derechos humanos y los principios democráticos.

Además, cabe destacar la mención que se hace en el Informe Moraes acerca de las motivaciones de las operaciones de vigilancia masiva, que distan de proteger únicamente a la seguridad nacional y combatir el terrorismo. Añade que son utilizadas especialmente para el espionaje económico industrial y para la elaboración de perfiles, tratando a todos los ciudadanos como potenciales sospechosos.

Destaca también el papel de las empresas privadas cuyos data centers hemos visto con anterioridad. El informe afirma que hay un elevado grado de implicación por su parte que se reflejan en varios factores provocados por las propias empresas: no establecer medidas de seguridad informática ni políticas de cifrado adecuadas, incorporar puertas traseras[34] a propósito y desobedecer la legislación en materia de protección de datos e intimidad.

A modo de resumen, estos serían los principales puntos relacionados con la vigilancia masiva que podemos extraer del Informe Moraes:

- Prueba la existencia de programas secretos tecnológicamente muy avanzados de vigilancia masiva.
- Vulneración de los derechos fundamentales de todos los ciudadanos.
- Motivaciones alejadas de la seguridad nacional y de la lucha anti-terrorista.
- Implicación de gran cantidad de Países miembros de la Unión Europea en la trama del espionaje de masas, además de los Estados Unidos.
- Implicación directa o indirecta de empresas de Internet y de telecomunicaciones.

[34] En informática, una puerta trasera o *backdoor* es una secuencia mediante la cual se pueden evitar los sistemas de seguridad de un algoritmo de autentificación para acceder a un sistema.

4.2 ¿De qué países y agencias hablamos?

En el Informe Moraes se nombran explícitamente los países que forman parte de las alianzas de los Nueve Ojos y de los Catorce Ojos, algo fundamental para comprender el alcance del espionaje masivo a nivel mundial y para determinar qué países europeos están directamente implicados en el uso de sistemas de vigilancia masiva.

La raíz principal de ambas alianzas se encuentra en la formación inicial, denominada alianza de los Cinco ojos. Para conocer los orígenes de este organismo, también conocido como "FVEY", debemos remontarnos a la década de los cuarenta, donde el clima bélico de la época hizo vital establecer una serie de alianzas estratégicas. Así pues, en 1943, ingleses y norteamericanos firmaron un primer acuerdo de cooperación entre sus servicios de inteligencia, conocido como ABRUSA. Tres años después se redactó un nuevo acuerdo, llamado AKUSA, que formó los primeros pilares de la colaboración entre la Agencia de Seguridad Nacional (NSA) norteamericana y el Cuartel General de Comunicaciones del Gobierno (GCHQ) inglés.

AKUSA se fue incrementando con la cooperación de nuevos países. Pese a que Noruega, Dinamarca y Alemania Occidental estuvieron implicadas en el acuerdo durante la década de los 50, fueron finalmente Canadá, Australia y Nueva Zelanda los países incluidos dentro del tratado, formando así los Cinco Ojos.

En definitiva, la alianza de los Cinco Ojos está formada por los servicios de inteligencia de señales (SIGINT) de Estados Unidos, Reino Unido, Canadá, Australia y Nueva Zelanda.

Tal y como se pudo ver en la documentación filtrada por Edward Snowden, si uno de esos países no dispone de una orden judicial para vigilar las comunicaciones de los ciudadanos de su propio país, se aplican con precisión las normas legales. Sin embargo, no tienen ese cuidado cuando se trata de ciudadanos extranjeros. De esa manera, si la NSA no tiene permiso para interceptar las comunicaciones de la ciudadanía norteamericana ya se encarga de ello la GCHQ inglesa, que compartirá todos esos datos con su homólogo americano.

Según fuentes como RT[35], los cinco países que forman los Cinco Ojos controlan alrededor del 90% del tráfico de comunicaciones de Internet, repartiéndose el control de las comunicaciones globales de la siguiente manera:

- EEUU: América Latina, Caribe, China, Rusia, Oriente Próximo y África
- Canadá: zonas del norte del Atlántico y del Pacífico, parte de Rusia y China
- Reino Unido: Europa y Rusia Occidental
- Australia: sur y este de Asia
- Nueva Zelanda: zona sur del Pacífico

La alianza de los Cinco Ojos cuenta además con los denominados 3rd Party Foreign Partners, una serie de países que comparten ocasionalmente con ellos sus labores de inteligencia y colaboran en proyectos conjuntos. Los países que forman parte de la 3rd Party se dividen en dos tipos: los que pertenecen al Tratado del Atlántico Norte y los aliados del Oeste.

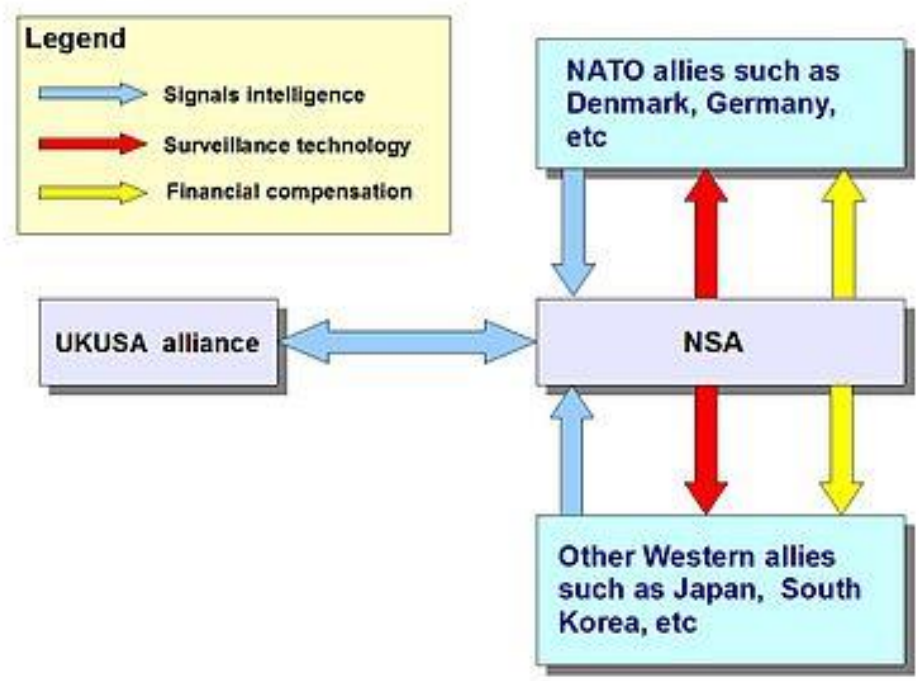

Tabla 17. Aliados de la NSA. (EdwardSnowden.com, 2015)

[35] RT. *El objetivo del 'Club de los Cinco Ojos' es "la supremacía económica sobre otros países"* [en línea]. RT, 2013 [Consulta: 15 mayo 2016] Disponible en: <https://actualidad.rt.com/actualidad/view/110072-club-cinco-ojos-supremacia-economica-espionaje>.

Entre los aliados del Oeste se encuentran países como Japón, Corea del Sur o Israel, y entre los aliados del Tratado del Atlántico Norte destacan tres clases de países. Por un lado, los que forman los Nueve Ojos, por otro, los que forman parte de los Catorce Ojos, y por último, otros países aliados europeos.

Los Nueve Ojos lo forman los cinco países que integran los Cinco Ojos más una serie de países que mantienen una relación de inteligencia relativamente estrecha con ellos: Dinamarca, Francia, Países Bajos y Noruega.

Los Catorce Ojos lo integran todos los miembros anteriores más Alemania, Bélgica, Italia, España y Suecia, y su propósito es coordinar el intercambio de señales militares y de inteligencia entre ellos.

Tabla 18. Clasificación de los países aliados de EEUU. (EdwardSnowden.com, 2015)

En definitiva, podemos determinar que, además de los Estados Unidos, los países democráticos de la Unión Europea implicados en un principio en el control y la vigilancia masiva ilegal de la ciudadanía global son: Alemania, Austria, Bélgica, Croacia, Dinamarca, España, Finlandia, Francia, Grecia, Hungría, Italia, Países Bajos, Polonia, Rumania, Reino Unido, República Checa y Suecia.

Por otro lado, si se revisa el análisis de la documentación filtrada por Snowden hecho por los medios de comunicación internacionales, se puede realizar un seguimiento de las agencias de inteligencia directamente implicadas que colaboran entre ellas en el intercambio de datos y, a veces, en el uso compartido de programas de vigilancia.

Países	Agencias de Inteligencia	Presupuesto anual
Alemania	Federal Intelligence Service (BND)	615,6 millones € (2015)[36]
Dinamarca	Danish Security and Intelligence Service (DSIS)	400 millones €[37]
España	Centro Nacional de Inteligencia Español (CNI)	255 millones € (Wikipedia)
EEUU	National Security Agency (NSA) Central Intelligence Agency (CIA) Federal Bureau of Investigation (FBI)	10,8 billones $ (2013) 14,7 billones $ (2013)[38] 8,7 billones $ (2016)[39]
Francia	General Directorate for External Security (DGSE)	731,8 millones € (Wikipedia)
Italia	External Intelligence and Security Agency (AISE)	Desconocido
Países Bajos	General Intelligence and Security Service (AIVD)	Desconocido
Reino Unido	Government Communications Headquarters (GCHQ)	1883 millones £ (Wikipedia)
Suecia	National Defence Radio Establishment (FRA)	860,2 millones SEK [91 millones €] (Wikipedia)

[36] Bundesministerium der Finanzen. *Einzelpläne* [en línea]. 2015 [Consulta: 13 mayo 2016]. Disponible en: <http://www.bundeshaushalt-info.de/#/2015/soll/ausgaben/einzelplan/0404.html>.
[37] Danish Security and Intelligence Service. *PETs finances* [en línea]. 2015 [Consulta: 13 mayo 2016]. Disponible en: <https://www.pet.dk/English/About%20PET/PETs%20finances.aspx>.
[38] Shane, Scott. *New leaked document outlines U.S. spending on intelligence agencies* [en línea]. Nueva York: The New York Times, 2013. [Consulta: 17 mayo 2016]. Disponible en: <http://www.nytimes.com/2013/08/30/us/politics/leaked-document-outlines-us-spending-on-intelligence.html?hp&pagewanted=all&_r=0>.
[39] FBI. *Mission & Priorities* [en línea]. U.S. Department of Justice, 2015 `Consulta: 17 mayo 2016]. Disponible en: <https://www.fbi.gov/about/mission>.

De todas las fuentes consultadas no he podido encontrar ninguna información que implique a los gobiernos de los siguientes Países Miembros: Austria, Bélgica, Croacia, Finlandia, Grecia, Hungría, Polonia, Rumania y República Checa. Por lo tanto, la lista definitiva de los países y agencias de inteligencia implicadas en el control y la vigilancia masiva son las de la tabla anterior.

4.3 Entrada y recopilación de datos

4.3.1 Datos susceptibles de interés

Determinar qué datos son susceptibles de interés para las agencias de inteligencia es una tarea relativamente sencilla si nos atenemos a las posturas que muestran al respecto agencias como la NSA o la GCHQ: todos.

Tal y como se puede ver en una presentación secreta que se llevó a cabo en la reunión anual de 2011 de la alianza de los Cinco Ojos, el lema es recogerlo todo en todo momento.

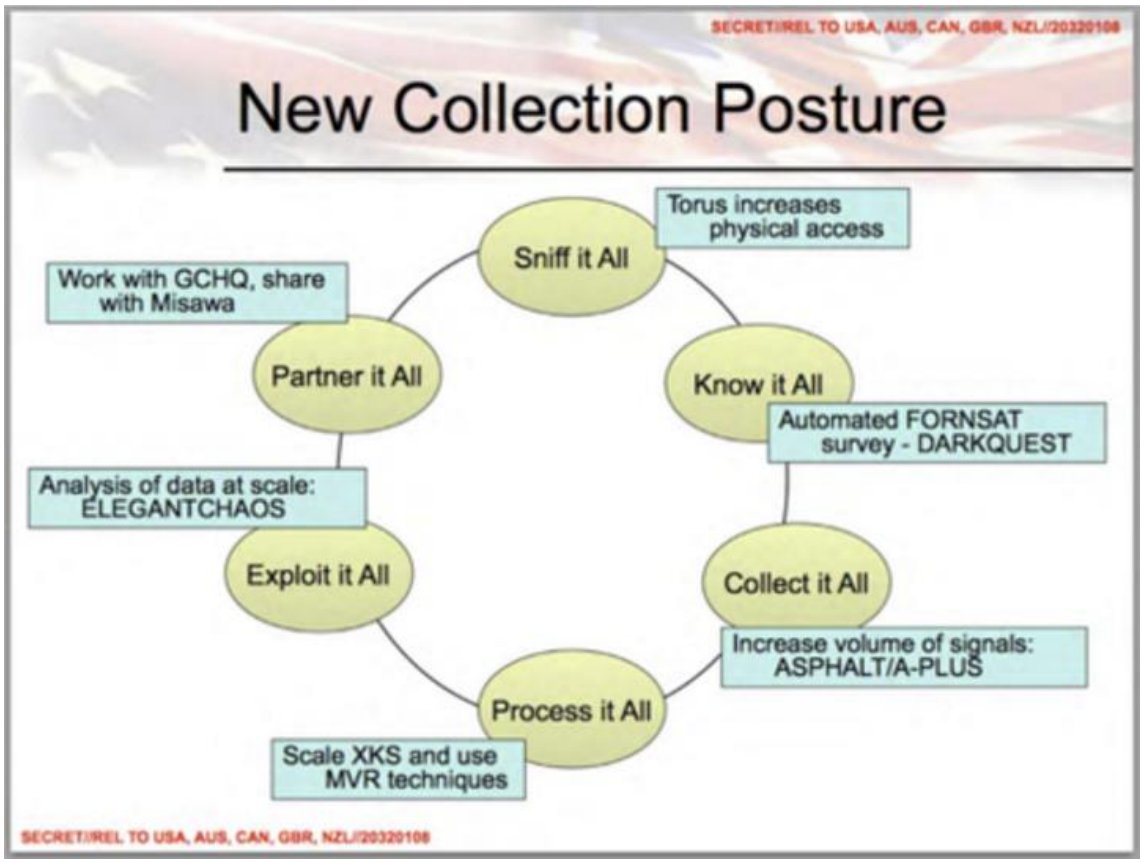

Tabla 19. Nueva postura de recolección en los Cinco Ojos. (EdwardSnowden.com, 2015)

¿Y qué es todo? Si observamos el tipo de datos que se recogen a través de los distintos programas de vigilancia masiva que han sido publicados en medios de comunicación de todo el mundo y los juntamos, podríamos determinar que "todo" es: historiales de búsquedas, contenido de correos electrónicos, transferencias de archivos, chats, fotografías, vídeos, videoconferencias, registros de conexiones, notificaciones de actividades, datos almacenados, datos de redes sociales, números de teléfonos, direcciones IP, datos de los navegadores utilizados, datos de geolocalización, registros médicos, datos económico, transacciones financieras, mensajes de texto…

Según el Wall Street Journal[40], solamente el sistema de interceptación de la NSA tenía en 2013 la capacidad para llegar aproximadamente al 75% de todo el tráfico de Internet de los Estados Unidos, una cifra que desde entonces muy probablemente haya crecido. Y recordemos que, tal y como he dicho con anterioridad, se estima que los países que forman parte de los Cinco ojos controlan el 90% del tráfico global de comunicaciones de Internet.

Contenido y metadatos

En términos generales, las agencias de inteligencia recogen dos tipos de información: contenido y metadatos.

- **Contenido:** llamadas telefónicas, e-mails, chats online, actividad general en Internet, historiales de navegadores y labores de búsqueda.
- **Metadatos:** datos sobre los datos anteriores. Es información sobre el contenido, pero no el contenido en sí.

Los metadatos sobre un e-mail registran, por ejemplo, el emisor y el destinatario del correo, el tema o la ubicación del remitente. En una llamada telefónica, los metadatos indican las identidades del que llama y del que recibe la llamada, la duración de la conversación, el emplazamiento o los tipos de aparatos usados.

[40] RTVE.es. *La NSA tiene capacidad para espiar el 75% del tráfico de Internet de Estados Unidos* [en línea]. Corporación RTVE, 2013 [Consulta: 22 mayo 2016]. Disponible en: <http://www.rtve.es/noticias/20130821/nsa-tiene-capacidad-para-espiar-75-del-trafico-internet-estados-unidos/741820.shtml>.

En un documento filtrado de la NSA nos podemos hacer una idea general de los metadatos que una agencia de inteligencia recopila y almacena de una llamada telefónica:

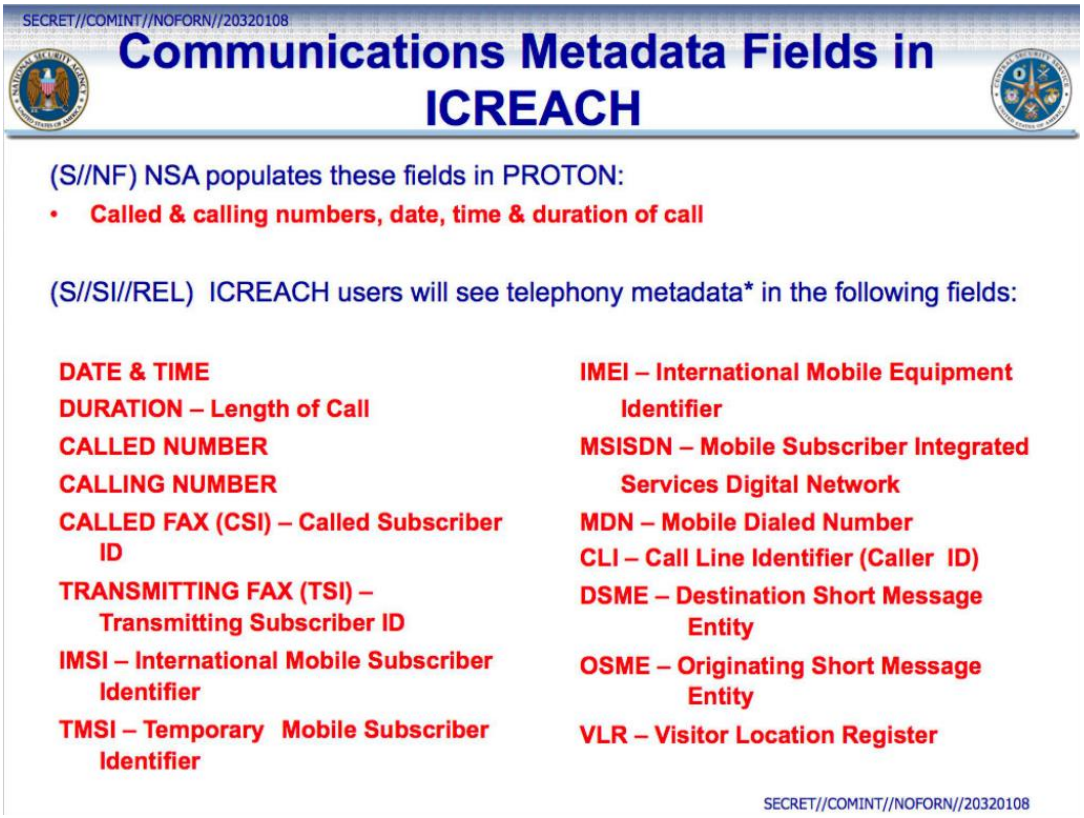

Tabla 20. Metadatos procedentes de llamadas telefónicas. (EdwardSnowden.com, 2015)

Tal y como podemos observar, los metadatos telefónicos que se recogen son los siguientes: fecha, tiempo y duración de la llamada, número del emisor y del receptor, la identidad internacional del abonado a un móvil (IMSI), la identidad temporal del abonado a un móvil (TMSI), la identidad internacional del equipo móvil (IMEI), la identidad del abonado integrado en la red digital de servicios integrados (MSISDN), los números marcados, el identificador de la línea de llamada o Caller ID (CLI), el destino y el origen de un mensaje corto y el registro de localización de visitante (VLR).

Para comprender mejor el alcance de lo que supone recopilar los metadatos anteriormente identificados, es necesario definir exactamente algunos de los conceptos anteriores.

Así pues, el **IMSI** es un código de identificación único para cada dispositivo de telefonía móvil que se encuentra integrado en la tarjeta SIM. Permite su identificación a través de las redes GSM y UMTS, que los sistemas de comunicaciones móviles de segunda y tercera

generación utilizan respectivamente. Es un código formado por el código del país (3 dígitos), el código de la red móvil (2-3 dígitos) y el número que identifica la estación móvil.

El **TMSI** es un código que identifica la identidad que se envía comúnmente entre el móvil y la red. La red asigna un código al azar para cada móvil en una zona en el momento en que éste se enciende. El número es local a un área de ubicación, por lo que se actualiza automáticamente cada vez que el móvil cambia a una nueva área geográfica. La red puede cambiar este código en cualquier momento, algo que suele hacer a menudo para evitar que el abonado sea identificado y rastreado, aunque como vemos, es una circunstancia salvable para las agencias de inteligencia como la NSA.

El **IMEI** es un código pregrabado en los móviles que utilizan la red global GSM (móviles de segunda generación). Identifica al dispositivo de forma exclusiva a nivel mundial y se transmite a la red cada vez que el móvil se conecta a ella. Informa acerca de la identidad del emisor, su localización y el terminal telefónico utilizado.

El **MSISDN** es el número de subscripción a la red digital de servicios integrados. Está formado por el código numérico del país, seguido del número de abonado a la red de teléfono. Una única tarjeta SIM puede permitir varios números MSISDN, lo que permite a los usuarios ser llamados desde distintos números con un único terminal.

El **CLI** es un servicio telefónico disponible en los sistemas de telefonía tanto analógicos como digitales y en la mayoría de las aplicaciones digitales de voz. Transmite el número de la persona que realiza la llamada al equipo telefónico del receptor antes de que se conteste la llamada. Permite también asociar un nombre al número de teléfono del emisor.

Finalmente, el **VLR** es una base de datos en una red de comunicaciones móviles que contiene la ubicación exacta de todos los abonados móviles presentes en un área de servicio de un Centro de Conmutación Móvil (MSC).

Por otro lado, en otra de las diapositivas filtradas podemos ver aún más la distinción entre contenido y metadato en función del tipo de dato.

UK SECRET STRAP1
COMINT AUS/CAN/NZ/UK/US EYES ONLY ORCON

CONTENT OR METADATA?

Categorisation of aspects of intercepted communications - GCHQ policy guidance.

POCs: [redacted] of GCHQ OPP-LEG [redacted]@gchq.gov.uk) and [redacted] of GCHQ OPP-HQ [redacted]@gchq.gov.uk).
Date last verified: 20 January 2010 by [redacted]. Version 2.

Data type sorted by data class (column B; defined below)	**Class** (see below)	**Content (c) or Metadata (m)**	**Note: For convenience, the terms "Metadata", "Events" and "Communications Data" are often used interchangeably, although only the last is defined in UK law (RIPA).**
attachment to e-mail, eg routing diagram, picture, video	a	c	
e-mail address in the body of a message	a	c	
name of a file attached to an e-mail	a	c	
authentication data to a communications service: login ID, userid, password	ac	m	unless sent inside the body of a communication
an e-mail inside a message	c	c	
bulletin board posting	c	c	
chat-room discussions	c	c	
content of a voice call	c	c	
content of an e-mail	c	c	
cookie as a whole (some elements are listed separately below)	c	c	
DTMF data, as opposed to dialling	c	c	
keystroke logs	c	c	
search results	c	c	
search strings	c	c	
SMS or IM text	c	c	
video	c	c	
voice mail boxes	c	c	
web cam transmissions	c	c	
web forms filled in by people	c	c	
'to' 'from', 'cc', 'bcc', and 'fwd' lines within e-mail header	ca	m	
chat aliases, chat handles and other related or similar identifiers	ca	m	
e-mail address from a cookie sent to set up a communication channel	ca	m	
IMEI data	ca	m	
IMSI caller ID	ca	m	
IP addresses of the computers sending and receiving the message	ca	m	
machine ID extracted from cookies (eg Yahoo B cookies)	ca	m	

UK SECRET STRAP1
COMINT AUS/CAN/NZ/UK/US EYES ONLY ORCON

Coding of data classes
a - attachment
ac - authentication of communications
c - content
ca - communications address
cdd - content derived data; some types in this class may be able to be treated as communications data due to low level of intrusion
ch - content header
d - traffic data including network management (excludes any such data sent inside the body of a communication, eg by a CSP)
m - miscellaneous; some types in this class may be able to be treated as communications data

bulk unselected: as taken from bearers without filtering or selection, save for national sensitivities

Data type sorted by data class (column B; defined below)	**Class** (see below)	**Content (c) or Metadata (m)**	**Note: For convenience, the terms "Metadata", "Events" and "Communications Data" are often used interchangeably, although only the last is defined in UK law (RIPA).**
personal IDs extracted from cookies, mail headers, chat sessions or buddy lists	ca	m	
telephone number	ca	m	
content summarization	cdd	cdd	
file type of an attachment to an email	cdd	cdd	
language or language fingerprinting	cdd	cdd	
speaker's gender	cdd	cdd	
speaker's ID	cdd	cdd	
e-mail subject line	ch	c	
dialling, signalling, routing, addressing or signalling information	d	m	
call charge records, including info about the time and length of a call	d	m	
Calling Line Identification, including numbers dialled	d	m	
details of routers and IP addresses that have handled the message	d	m	
DTMF dialling (not data)	d	m	
location of parties to a communication, not derived from content	d	m	
logs of visitors to chat rooms including how often they have visited/posted	d	m	
network management, eg billing, authentication or tracking of communicants	d	m	
start and finish time of an internet session or phone call	d	m	
status of chat sites, ie whether they are active and how many participants	d	m	
URLs up to and including the domain name	d	m	
session initiation protocols	d	m	
website registers including owner details; assume not UKUSA owned/registered	d	m	
creation of/access to draft message	m	m	N.B. the contents of a draft message are Content.
history of websites browsed (full URLs)	m	c	
buddy lists for web mail, instant messenger or social networking	m	m	
folders used to organise e-mails	m	m	
address books or contacts lists for web mail etc	m	c	
crypto keys	m	c	
password to internet or telephony services other than communications services	m	c	
URLs beyond the domain name, ie one that may include search terms	m	c	

Tabla 21. ¿Contenido o metadato? Guía de categorización. (EdwardSnowden.com, 2015)

Así pues, a modo de dejemplo, los resultados de una búsqueda serían datos de contenido, y la IP de un ordenador que manda o recibe un mensaje sería un metadato.

Cuando saltó el escándalo por las filtraciones de documentos secretos por parte de Edward Snowden, el gobierno de los Estados Unidos insistió vehementemente en que buena parte de la vigilancia revelada en la colección de Snowden se refiere a la recogida de metadatos, no de contenido, como dando a entender que es un tipo de espionaje menos intrusivo que uno que se dedica a recopilar más cantidad de contenido. La realidad, sin embargo, es que la recopilación del tipo de metadatos que hemos visto en las figuras anteriores es más intrusiva que la interceptación de contenidos.

Para hacernos una idea, en el libro de Glenn Greenwald *Sin un lugar donde esconderse*, así exponía el profesor de ciencia informática de Princeton, Edward Felten (2013), el por qué la vigilancia de metadatos puede ser especialmente significativa:

"*Veamos el siguiente ejemplo hipotético: una mujer joven llama a su ginecólogo; a continuación llama a su madre y luego a un hombre al que en los últimos meses ha telefoneado una y otra vez a partir de las once de la noche; a eso sigue una llamada a un centro de planificación familiar donde también se realizan abortos. Surge un probable guion que no habría sido tan evidente si solo se hubiera examinado el registro de una única llamada*".

En una llamada telefónica los metadatos suelen ser bastante más informativos que el contenido de la misma. De hecho, las escuchas de llamadas pueden resultar bastante difíciles debido a diferencias lingüísticas, el uso de argot o códigos o cualquier cosa que confunda el significado. Y ya no digamos analizar ese cúmulo desestructurado de manera automatizada. Los metadatos son nítidos y precisos, y lo más importante, son fáciles de analizar. Los metadatos no están sujetos a esas restricciones y pueden informar sobre muchísimas cosas relativas a costumbres, asociaciones, patrones de comportamiento, rutinas, hábitos, afiliaciones, aptitudes sociales…

La recogida masiva de metadatos permite no solo obtener información sobre más personas, sino también enterarse de hechos nuevos, antes privados, de los que no se habría sabido nada si solamente el espionaje se limitara a la recogida de contenido.

4.3.2 Mecanismos de recopilación

Para reunir semejante cantidad de datos – "todo" – la agencias de inteligencia se basan en múltiples métodos, como acceder directamente a cables internacionales de fibra óptica (incluidos los submarinos), desviar mensajes a depósitos cuando atraviesan los sistemas de comunicaciones de sus países, cooperar con otros servicios de inteligencia y contar con la información recogida de los clientes de las grandes empresas de Internet y de telecomunicaciones, ya sea voluntariamente o no.

Para hacerlo más comprensible, podríamos dividir la forma en que las agencias de inteligencia recopilan datos en cuatro grandes mecanismos: mediante la interceptación, mediante la cesión, mediante la compra y mediante la colaboración.

La recopilación de datos a través de la interceptación hace referencia al tipo de vigilancia secreta que ejercen los gobiernos y sus agencias de inteligencia a través de mecanismos tecnológicos capaces de interceptar, recolectar, almacenar y analizar grandes conjuntos de datos procedentes de cualquier fuente, normalmente procedentes de los data centers de las grandes empresas tecnológicas y de los cables submarinos de fibra óptica. Son los denominados Programas de Vigilancia Masiva (PVM), entre los cuales destaca especialmente PRISM, famoso por el trato de la prensa y del que hemos hablado superficialmente con anterioridad a la hora de determinar las empresas a estudiar.

Por otro lado, la recopilación de datos mediante la cesión hace referencia a todos aquellos datos que ceden a las agencias voluntariamente o por obligación los grandes data centers de las empresas vistas en el anterior apartado y de las grandes compañías de telecomunicaciones. En este sentido, destaca especialmente el caso de la compañía de telecomunicaciones Verizon, obligada a través de una orden judicial a entregar a la NSA todos los datos relacionados con todos sus clientes norteamericanos.

La recopilación de datos a través de la compra hace referencia a un número extraordinario de empresas privadas que venden software especializado en ciberseguridad y se dedican a desarrollar y vender troyanos, virus, antivirus, spammers, malwares, spywares y softwares de monitoreo ilegal de redes. Son empresas que trabajan codo con codo con gobiernos de todo el mundo.

Finalmente, cabe destacar que gran parte de la recopilación y entrada de datos a los sistemas de las diversas agencias de inteligencia no sería posible sin la colaboración con otras agencias. Hay un considerable grueso de datos que son fruto de la compartición de información entre agencias, información a la que por sí mismas no pueden acceder por diversos motivos. Es en este caso donde entran en juego las alianzas vistas con anterioridad de los Cinco, Nueve y Catorce Ojos.

Interceptación

Teniendo en cuenta la información sobre los Programas de Vigilancia Masiva (PVM) divulgada en los medios de comunicación de todo el mundo a partir de las filtraciones de Edward Snowden y del informe Moraes, podemos determinar que los principales y más destacables PVMs utilizados en la interceptación de datos privados son los siguientes: PRISM, XKeyScore, TEMPORA, Bullrun, Edgehill, Quantum Inside, Dishfire y Muscular.

Cabe destacar que son programas extremadamente complejos que no se dedican exclusivamente a recopilar datos, ya que la gran mayoría son capaces de realizar el análisis de los datos que recopilan, incluirlos en sus propias bases de datos y facilitar la recuperación de los datos a través de sus propios sistemas de búsqueda y recuperación.

En el citado Informe Moraes se reconoce la existencia de los PVM y los define de la siguiente manera: "*sistemas tecnológicamente muy avanzados, complejos y de amplio alcance diseñados por los servicios de inteligencia de los Estados Unidos y de algunos Estados miembros para recopilar, almacenar y analizar datos de comunicaciones, incluidos datos de contenido y datos y metadatos de localización de todos los ciudadanos en todo el mundo a una escala sin precedentes y de una manera indiscriminada y no basada en sospechas*".

PRISM

Sin duda, uno de los PVM más conocidos por el trato que tuvo en la prensa mundial, es PRISM. Se trata de un programa creado por la NSA que se encuentra operativo desde 2007. PRISM destaca por su capacidad de interceptar información mediante el acceso directo a los servidores centrales de empresas estadounidenses líderes en internet, como Google, Microsoft, Facebook, Yahoo, Skype o Apple.

PRISM puede acceder a una amplia gama de comunicaciones directamente desde los servidores de las principales empresas de Internet: historiales de búsquedas, contenidos de correos electrónicos, transferencias de archivos, chats, fotografías, videoconferencias, registros de conexiones…

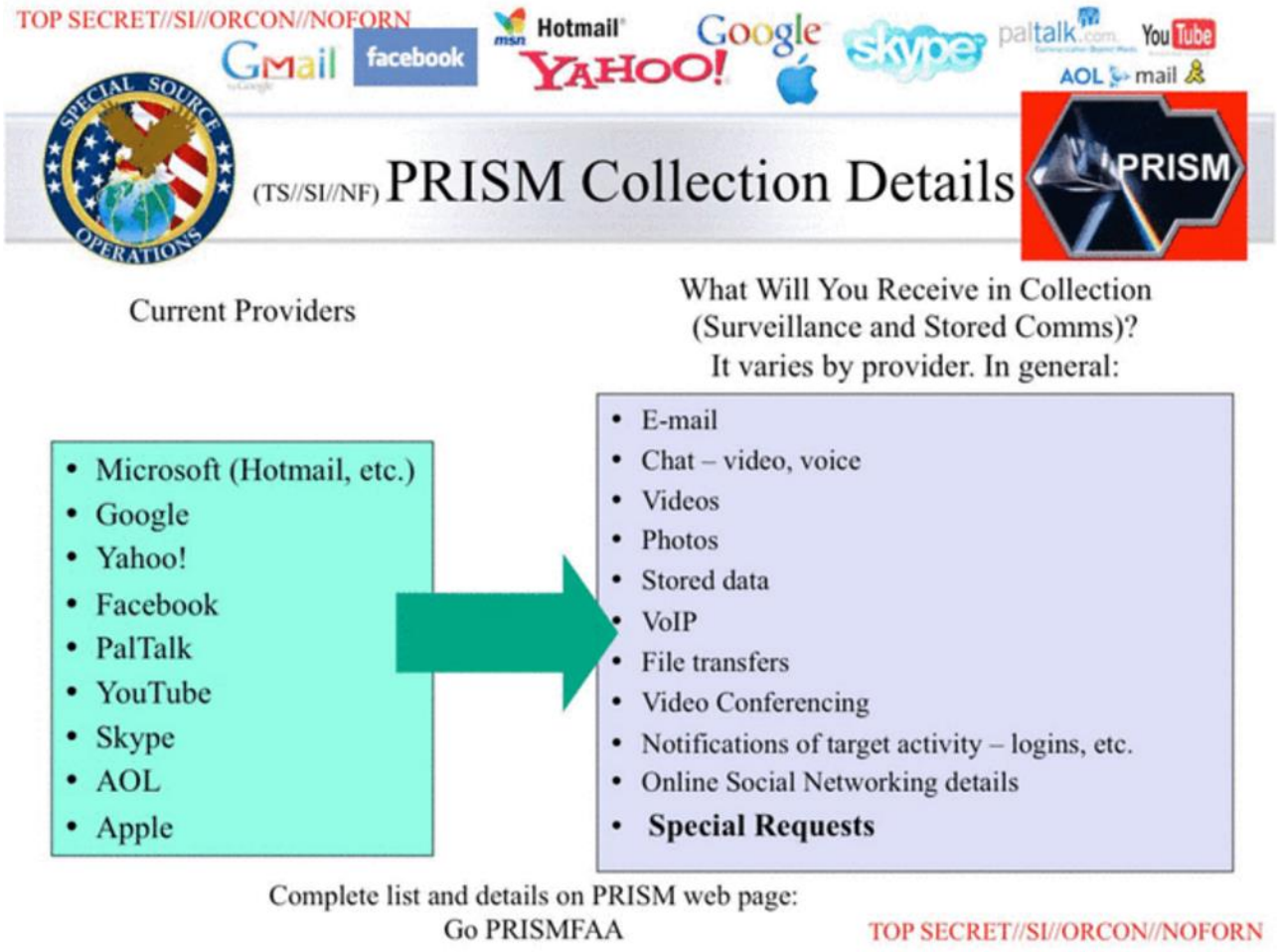

Tabla 22. PRISM, detalles de proveedores y datos recibidos. (EdwardSnowden.com, 2015)

PRISM cuenta con las siguientes particularidades técnicas, tal y como podemos ver en la figura inferior:

- Dispone de 9 proveedores de servicios con base en EEUU que dan acceso a los denominados selectores DNI (Digital Network Intelligence), es decir, a cualquier actividad realizada vía Internet.
- No dispone de proveedores de selectores DNR (Dialed Number Recognition), es decir, datos telefónicos; o al menos no disponía de ello en el momento en que se realizó la diapositiva.
- Puede acceder al contenido de las comunicaciones almacenadas, es decir, permite realizar búsquedas de comunicaciones concretas.
- Es capaz de procesar datos a tiempo real
- Es capaz de recolectar datos de voz.
- No puede recoger datos de geolocalización

Tabla 23. Características de PRISM. (EdwardSnowden.com, 2015)

Por otro lado, el mecanismo estructural de PRISM permite a la NSA y al FBI llevar a cabo la vigilancia en tiempo real de correos electrónicos y mensajería instantánea, aunque todavía no está claro cuáles son los proveedores específicos de servicios de Internet permiten que este tipo de vigilancia.

El funcionamiento es tan concreto y la vigilancia tan específica, que la NSA recibe notificaciones de cuándo el objetivo - una persona vigilada - inicia sesión a un determinado servicio, envía un correo, envía un mensaje de texto, voz o realiza una sesión de chat de voz o texto.

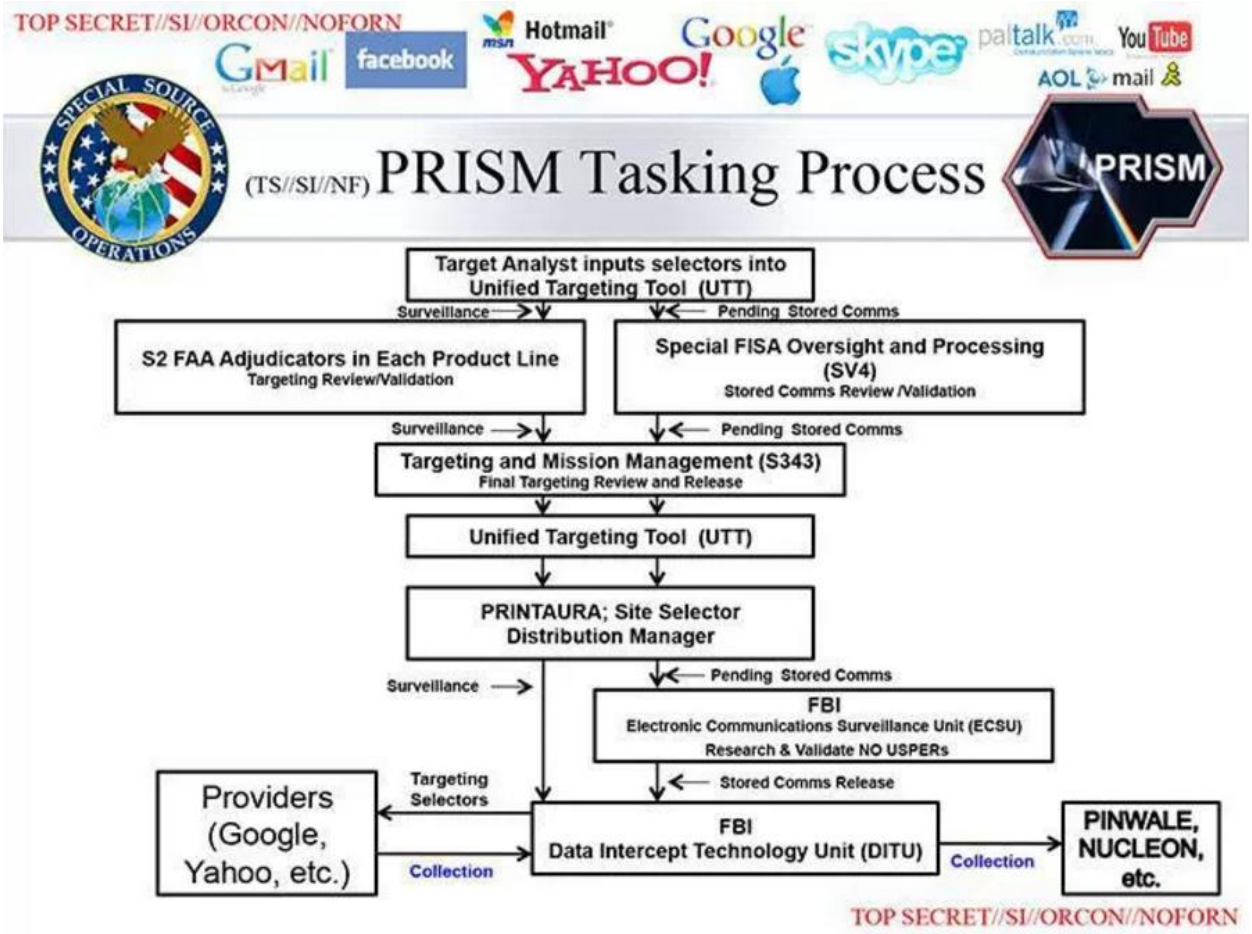

Tabla 24. PRISM, cadena de mando. (EdwardSnowden.com, 2015)

En la anterior figura se describe lo que sucede cuando un analista de la NSA identifica a un objetivo para vigilar y le pasa al sistema PRISM esa información. Esa información pasa a manos de un supervisor, que se encarga de verificar que el objetivo en cuestión no es un ciudadano estadounidense o un extranjero en suelo norteamericano. El supervisor da su visto bueno en base a una sospecha razonable y da vía libre.

Durante este proceso, parece ser que es el FBI quien se encarga de usar la tecnología de PRISM para acceder directamente a los servidores de las compañías que participan en el programa, que recordemos, son Google, Facebook, Microsoft, Yahoo, Apple, PalTalk y AOL.

Una vez interceptados los datos de los servidores de estas compañías, estos pasan directamente a la NSA, a la CIA y al propio FBI sin ningún tipo de revisión.

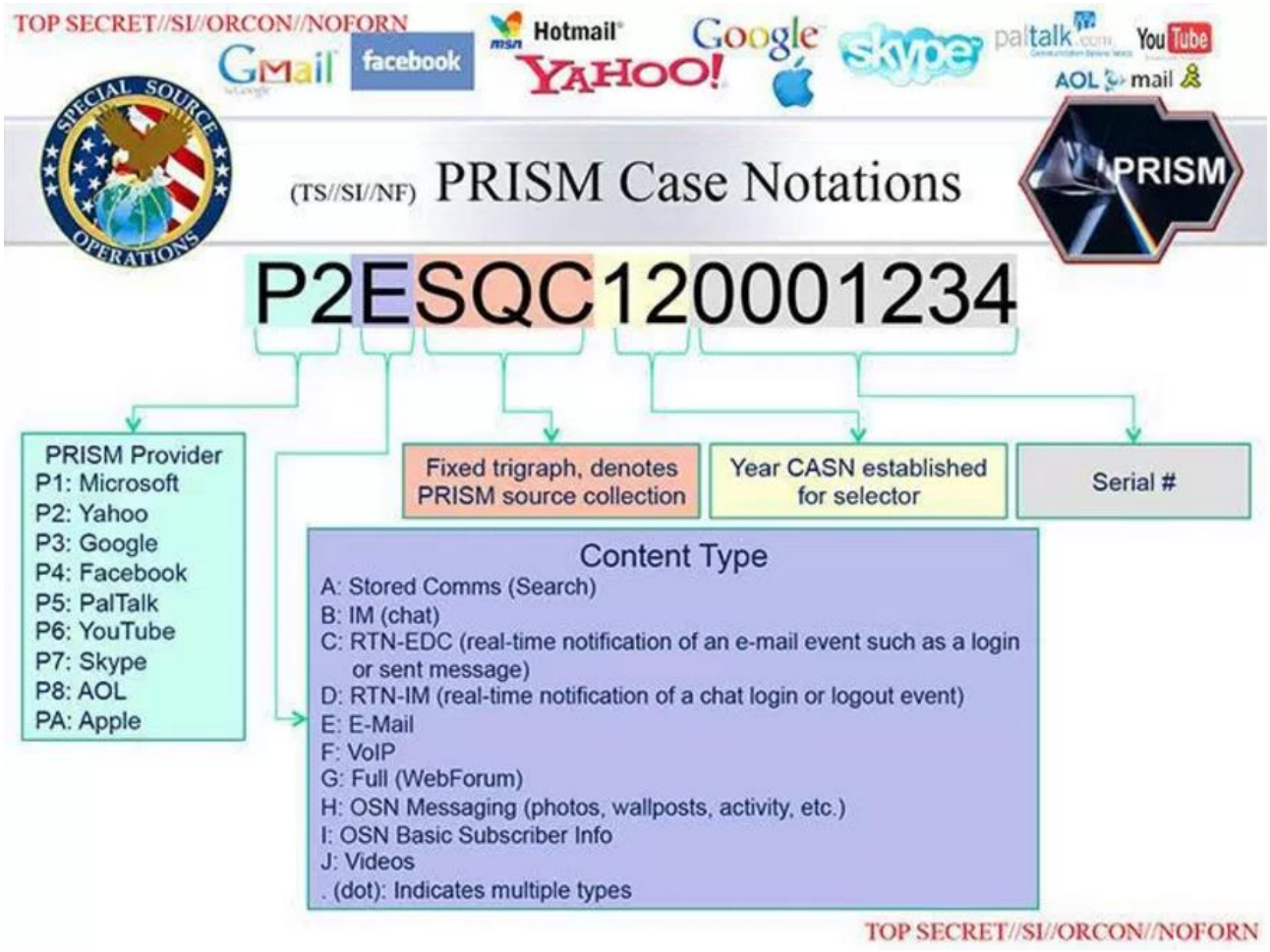

Tabla 25. PRISM, ejemplos de casos. (EdwardSnowden.com, 2015)

Finalmente, en la anterior figura se puede observar que cada objetivo tiene un identificador único compuesto por el proveedor de origen de los datos, el tipo de contenido y códigos automáticos establecidos por el sistema. Destaca el apartado de Content Type, donde se ve cómo PRISM permite la vigilancia a tiempo real de mensajes, correos electrónicos y chats.

XKeyScore

A pesar de que PRISM es el PVM que más impacto ha tenido en los medios de comunicación globales, el programa XKeyScore de la NSA es, sin lugar a dudas, más intrusivo en cuanto a privacidad ciudadana respecta. XKeyScore no es solamente un sistema de interceptación y recolección de datos, es un sistema formado por múltiples interfaces, bases de datos, servidores y recolectores de metadatos que entran al sistema a través de muchos otros programas. Está formado por unos 700 servidores localizados en 150 sitios alrededor del mundo.

En relación a toda la información que entra al sistema de XKeyScore a través de otros sistemas, podemos ver que la estructura del programa es la siguiente:

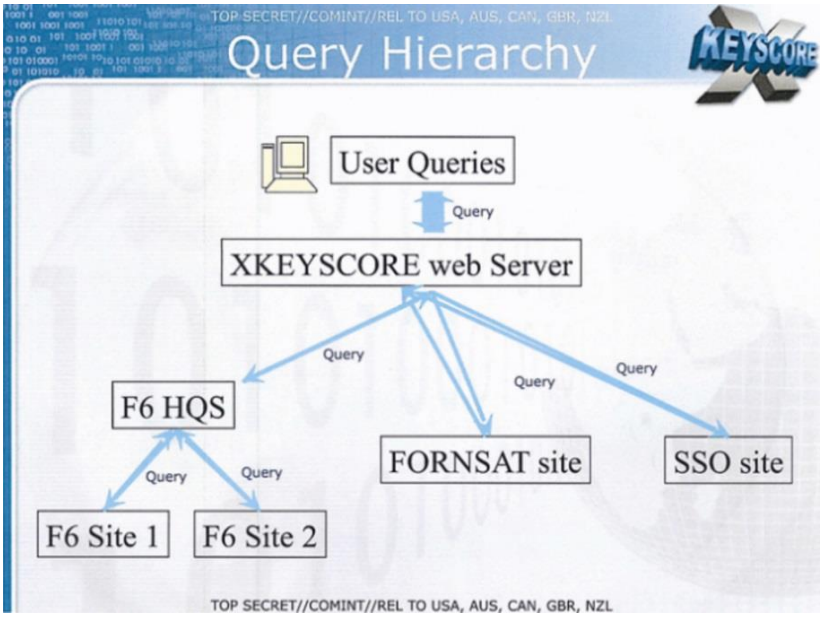

Tabla 26. Jerarquía en XKeyScore. (EdwardSnowden.com, 2015)

De esta manera, los servidores de XKeyScore, que almacenan los datos procedentes de otros programas, también se engrosan con la información que proviene de los siguientes sistemas de recolección:

- **F6HQS (Special Collection Service):** espionaje de diplomáticos y líderes políticos extranjeros.
- **FORNSAT (Foreign Satellite Collection):** interceptación de datos de satélites extranjeros.
- **SSO (Special Source Operations):** recopilación de datos procedentes de proveedores de servicios de telecomunicaciones (Vodafone, Verizon, etc.)

Además de estos sistemas, en otra diapositiva se mencionan más fuentes de recolección de datos:

Tabla 27. Fuentes de información del programa XKeyScore. (EdwardSnowden.com, 2015)

- **Overhead:** datos procedentes de aviones espía, drones y satélites norteamericanos.
- **Tailored Access:** datos procedentes de operaciones relacionadas con hackers y guerras cibernéticas.
- **FISA:** todos los datos que proceden de la vigilancia permitida y aprobada por el Tribunal de Vigilancia de Inteligencia Extranjera de los Estados Unidos.
- **3rd Party:** datos recopilados por las agencias de inteligencia consideradas como socios extranjeros de la NSA

Por otro lado, tal y como podemos ver en el artículo que Gleen Greenwald escribió en el The Guardian[41] el 31 de julio de 2013, basándose en el material revelado por Edward Snowden, XKeyScore es el sistema de mayor alcance de la NSA en cuanto a la red de inteligencia digital, o tal y como lo denomina la propia agencia, de DNI, Digital Network Intelligence.

[41] https://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data

El propósito de XKeyScore es dar acceso a los analistas a los metadatos y al contenido de correos electrónicos y de cualquier actividad llevada a cabo en Internet, incluso sin haber una cuenta de correo electrónico conocida asociada al objetivo, ya que los analistas pueden buscar por nombre, número de teléfono o móvil, dirección IP, palabras clave, tipo de navegador utilizado e incluso por el idioma en que se ha llevado a cabo la actividad en Internet.

Lo que la NSA denomina como "selector fuerte" hace referencia a datos sumamente identificativos, como la dirección de correo electrónico, el número de teléfono o la dirección IP. Los "selectores suaves" serían el resto de datos no tan identificativos.

Es importante tener esto en cuenta, porque si la búsqueda solo fuera por correo electrónico (un selector fuerte) sería extremadamente limitada, ya que gran parte de las acciones llevadas a cabo en Internet son anónimas y no se pueden encontrar usando solamente direcciones de correo electrónico. Buscando a través de "selectores suaves", el analista puede encontrar el contenido deseado y extraer de él un selector fuerte, además de detectar muchísima información que habría sido imposible de encontrar usando solamente selectores fuertes. Esta es la gracia de XKeyScore. La mayor parte de los programas de vigilancia de la NSA operan a través de selectores fuertes, lo que genera volúmenes de datos en bruto demasiado elevados.

Entre los diversos campos de información del programa destacan: cada dirección de correo electrónico vista en una sesión por nombre de usuario y/o por dominio, cada número de teléfono visto en una sesión y cualquier cosa relacionada con la actividad del usuario, el correo electrónico, conversaciones de chats, nombres de usuario, listas de amigos o sistemas de cookies[42].

En la siguiente figura se puede observar cómo es el sistema de recuperación de correos electrónicos de XKeyScore. El analista solamente tiene que introducir la dirección de correo electrónico del objetivo, una justificación de su búsqueda y escoger el periodo de tiempo deseado.

[42] Información enviada por un sitio web y almacenada en el navegador del usuario, de manera que el sitio web puede consultar la actividad previa del usuario.

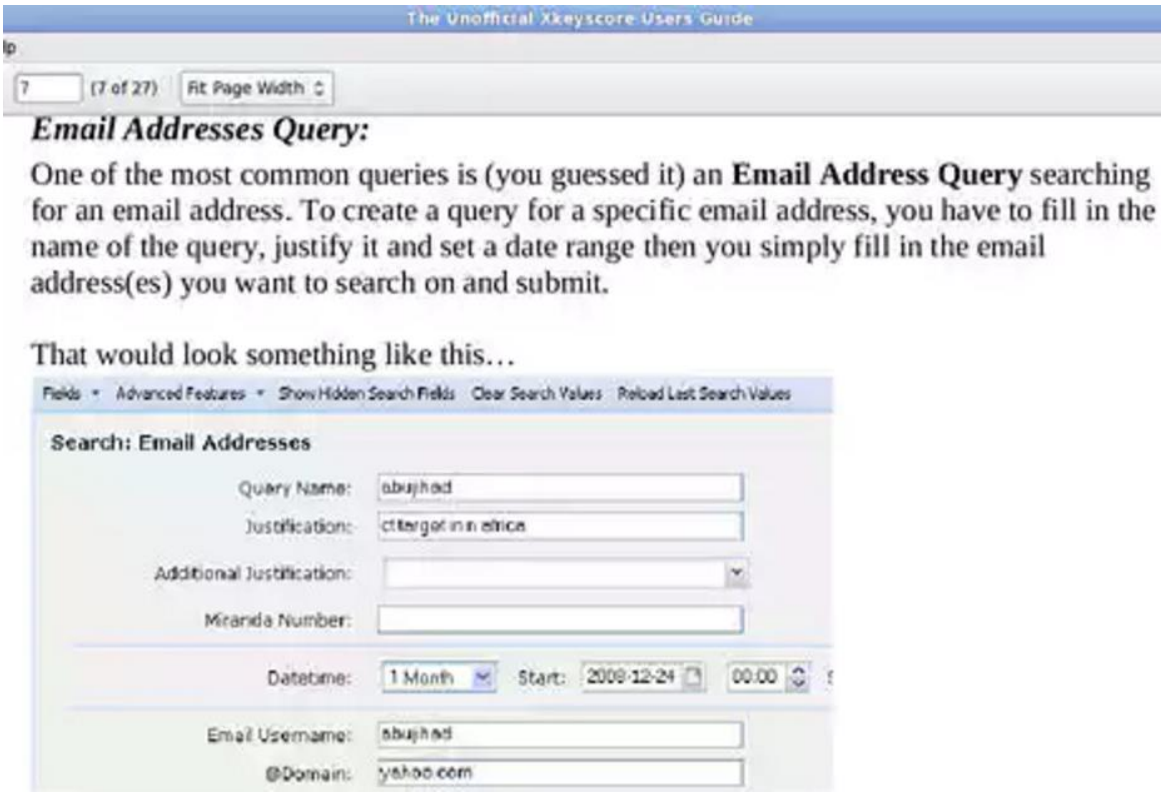

Tabla 28. Interfaz de búsqueda de correos electrónicos de XKeyScore. (EdwardSnowden.com, 2015)

A partir de aquí, en la interfaz de resultados, el analista debe seleccionar cuál de los e-mails recuperados por el sistema desea leer a través del software de lectura de la NSA.

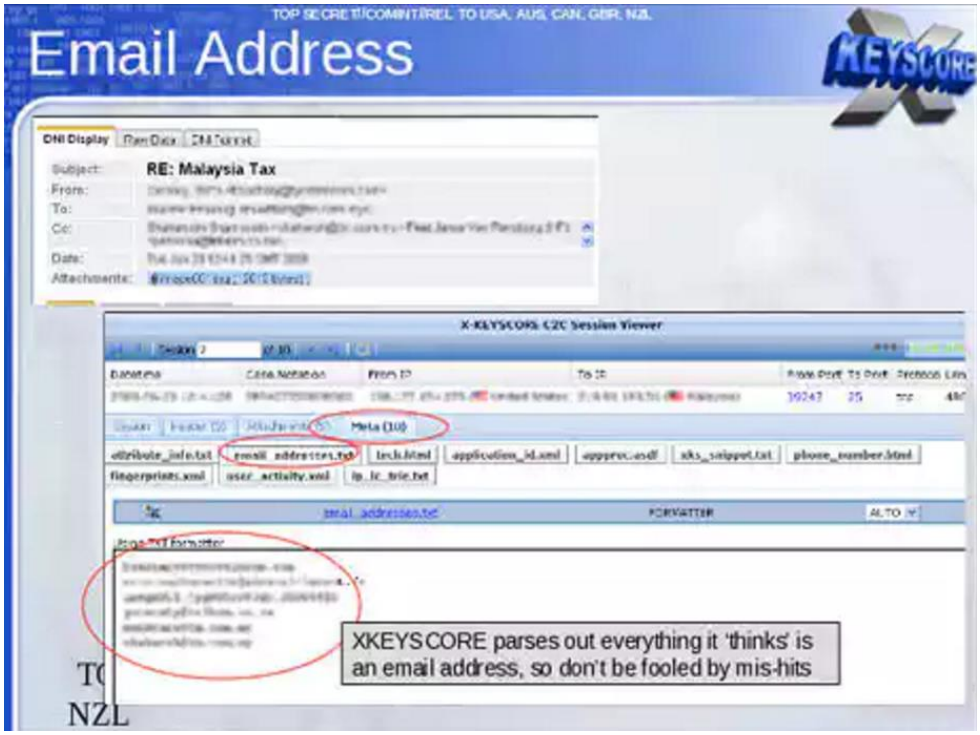

Tabla 29. Interfaz de resultados de correos electrónicos de XKeyScore. (EdwardSnowden.com, 2015)

Además de los correos electrónicos, tal y como he mencionado con anterioridad, XKeyScore es capaz de interceptar y registrar los datos de cualquier individuo que realice una actividad en Internet.

Un ejemplo del registro de actividades en Internet lo tenemos en una herramienta denominada DNI Presenter, que además de permitir acceder al contenido y a los metadatos de cualquier correo electrónico almacenado, también permite leer el contenido de los mensajes privados y chats de Facebook. Solamente es necesario que el analista introduzca en la base de datos del programa el nombre de usuario del objetivo y un intervalo de fechas en una pantalla de búsqueda simple.

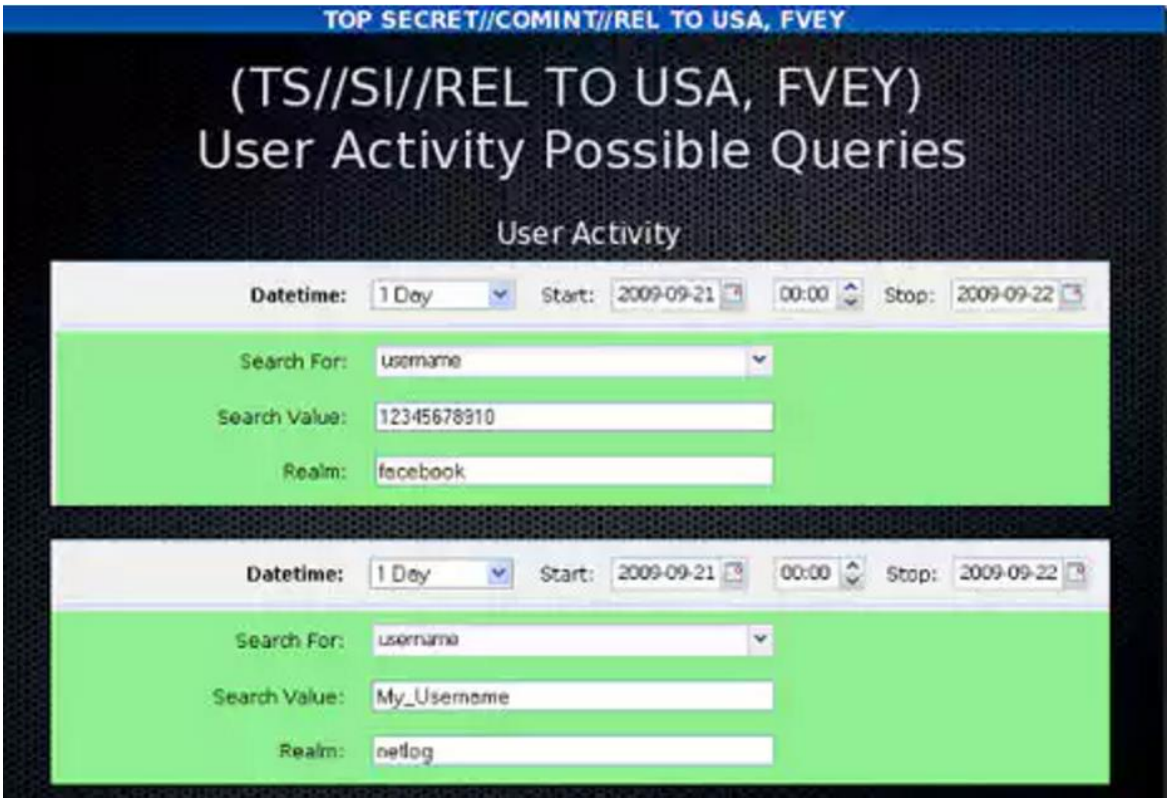

Tabla 30. Interfaz de DNI Presenter de XKeyScore. (EdwardSnowden.com, 2015)

Otro ejemplo del control de Internet lo tenemos en las actividades de navegación. Los analistas pueden buscar a través de una amplia gama de datos, incluyendo los términos de búsqueda introducidos por el usuario o las páginas web visitadas.

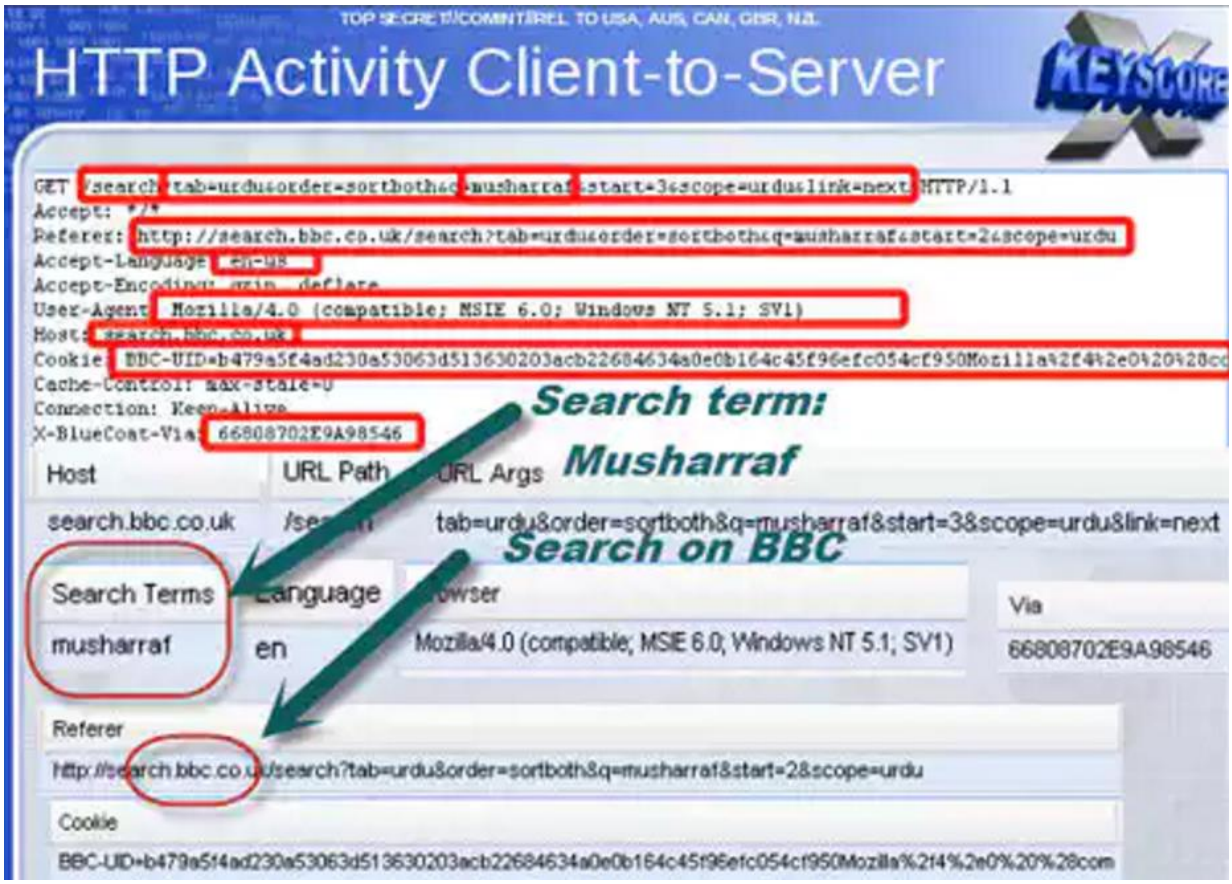

Tabla 31. Interfaz de búsqueda de actividades de navegación de XKeyScore. (EdwardSnowden.com, 2015)

La cantidad de comunicaciones accesibles a través de XKeyScore es gigantesca y abrumadora. Para hacernos una idea, un informe de la NSA en el 2007 estimó que por aquel entonces se habían recogido y almacenado más de 850 billones de registros de llamadas y cerca de 150 billones de registros de Internet, introduciendo de media de 1 a 2 billones de nuevos registros al día. Hablamos de cifras que datan de 2007, lo que quiere decir que actualmente, nueve años después, gracias a los avances tecnológicos en el campo de la informática y la ingeniería de sistemas, estas cifras se habrán incrementado hasta límites insospechados.

De hecho, durante el 2012, la agencia había registrado alrededor de 20 trillones de transacciones telefónicas y de e-mails solamente de ciudadanos estadounidenses. Y en un solo mes, XKeyScore tenía, por lo menos, 41 billones de registros recogidos y almacenados por un periodo de 30 días.

En cuanto al sistema de almacenaje del programa, es interesante destacar el hecho de que recoge tanta información de tantas fuentes diferentes que solamente se puede almacenar durante cortos periodos de tiempo. De esta manera, los datos de contenido permanecen almacenados en el sistema de tres a cinco días, mientras que los metadatos recolectados pueden guardarse hasta 30 días. Para solventar este inconveniente, la NSA almacena los datos en otras bases de datos, a parte de la propia base de datos de XKeyScore.

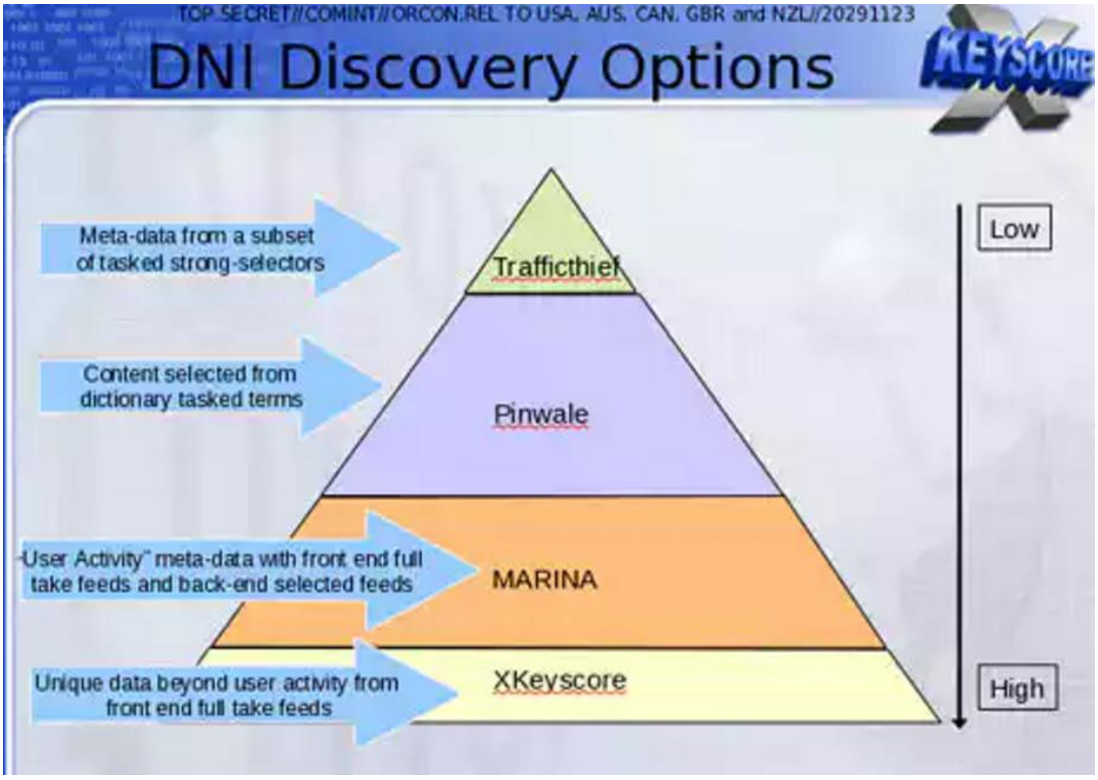

Tabla 32. Bases de datos donde se alojan los datos de XKeySocre. (EdwardSnowden.com, 2015)

A modo de resumen, las capacidades de XKEYSCORE las podemos ver reflejadas en una declaración de Edward Snowden que hizo durante una entrevista[43] el 26 de enero de 2014 para un canal de televisión alemán. A la pregunta del periodista "¿qué podría hacer si pudiera usar XKEYSCORE?", Snowden respondió:

"*Podría leer el correo electrónico de cualquier persona en el mundo de la que tuviera su dirección electrónica. Cualquier web: ver el tráfico que entra y sale. Cualquier ordenador en el que se siente alguien: puedes verlo. Cualquier portátil que quieras seguir: puedes seguirlo se mueva por donde se mueva. Puedes etiquetar a individuos… Por ejemplo, si usted trabaja en una gran empresa alemana y yo quiero acceder a su red, puedo rastrear su nombre de usuario a través de una página web o de un formulario en alguna parte, rastrear su nombre real, rastrear asociaciones con su red de amigos y construir lo que se denomina huella dactilar, que es la actividad en la red asociada a usted y que es única, lo que significa que conoceré su identidad esté donde esté aunque intente ocultar su presencia on-line. Es su identidad*".

[43] NDR.de. *Snowden-Interview: transcript* [en línea]. Norddeutscher Rundfunk, 2014 [Consulta: 4 junio 2016]. Disponible en: <https://web.archive.org/web/20140128224400/http://www.ndr.de/ratgeber/netzwelt/snowden277_page-1.html>.

TEMPORA

El PVM TEMPORA queda definido perfectamente en uno de los documentos filtrados por Edward Snowden. El documento en cuestión[44] es un pdf elaborado en septiembre de 2012 por la NSA. Es como una versión mejorada de XKeyScore.

El programa fue creado por la agencia británica GCHQ, y tal y como se puede observar en el documento, la NSA también puede gestionar y acceder al programa mediante la interfaz central de XKeyScore. La frase que introduce el documento es muy significativa: "todos hemos oído hablar de Big Data; ahora puedes tener un Gran Acceso a Big Data [Big Access to Big Data]".

TEMPORA es 10 veces más grande que XKeyScore. Utiliza alrededor de 10 mil máquinas para procesar y hacer posible a los analistas el acceso a más de 40 billones de registros al día. A pesar de tener mucha más capacidad, en TEMPORA también encontramos el límite de tres días de almacenaje para contenido y de 30 días para metadatos.

La capacidad de interceptación de datos también es superior a la de su homólogo norteamericano, ya que se nutre principalmente del tráfico telefónico y de todo el tráfico de datos que pasa por cables de fibra óptica de diferentes países. Así pues, TEMPORA, al igual que XKeyScore, permite interceptar grabaciones enteras de llamadas, metadatos, el contenido de correos electrónicos, el número de entradas de Facebook, el historial de acceso a páginas web de cualquier usuario de Internet… pero a mayor nivel.

De hecho, según un artículo publicado el 21 de junio de 2013 en el The Guardian[45], gracias a TEMPORA la inteligencia británica genera mucha más cantidad de colecciones de metadatos que la NSA debido a su más amplio acceso a los cables de fibra óptica.

---

[44] EdwardSnowden.com. *Tempora: the world's largest XKeyScore* [en línea]. 2012 [Consulta: 10 junio 2016]. Disponible en: <https://search.edwardsnowden.com/docs/TEMPORA%E2%80%94%E2%80%9CTheWorld%E2%80%99sLargestXKEYSCORE%E2%80%9D%E2%80%94IsNowAvailabletoQualifiedNSAUsers2014-06-18nsadocs>.

[45] MacAskill, Ewen [èt.al]. *Mastering the Internet: how GCHQ set out to spy on the world wide web* [en línea]. The Guardian, 2013 [Consulta: 6 junio 2016]. Disponible en: <https://www.theguardian.com/uk/2013/jun/21/gchq-mastering-the-internet>.

Tal y como expresa el mismo artículo, la existencia de un PVM como TEMPORA puede ser altamente peligrosa para la privacidad de cualquier ciudadano en el mundo que envíe o reciba e-mails, haga o reciba una llamada telefónica o ponga un mensaje en alguna red social, ya que la recopilación de datos que lleva a cabo se practica sin ningún fundamento de sospecha previo (no discrimina objetivos) y se ejecuta sin la necesidad de que se imponga alguna orden o permiso gubernamental.

En otro artículo del mismo periódico[46] publicado el mismo día, podemos ver una serie de cifras muy significativas al respecto. De acuerdo con el artículo en cuestión, en el año 2010 habría habido un total de 850 mil empleados repartidos entre la GCHQ, la NSA y contratistas privados con acceso a TEMPORA. Además, durante ese mismo año, TEMPORA fue capaz de interceptar 600 millones de registros telefónicos al día gracias al acceso de la inteligencia británica a más de 200 cables de fibra óptica y a la capacidad del programa de procesar los datos de 46 de estos cables a la vez.

En ese año, cada cable de fibra óptica transportaba datos a una velocidad de 10GB por segundo, por lo que, en teoría, TEMPORA tenía la capacidad de recoger más de 21PB al día. O lo que es lo mismo, sería como enviar toda la información que contienen todos los libros de la British Library 192 veces cada 24 horas.

Hablamos de 6 años atrás, por lo que conviene destacar que, evidentemente, la capacidad de TEMPORA habrá aumentado considerablemente a día de hoy, sobre todo teniendo en cuenta que, según el artículo, los documentos a los que el periódico ha tenido acceso dejan claro el esfuerzo constante de la GCHQ por aumentar su capacidad de recogida y almacenamiento, establecer acuerdos con más compañías de telecomunicaciones y ampliar sus capacidades técnicas a medida que los cables de fibra óptica transporten más datos a mayor velocidad.

Según el artículo, toda la intervención realizada por la agencia británica sobre los cables de fibra óptica a lo largo de cinco años no habría sido posible sin los acuerdos establecidos con algunas compañías de telecomunicaciones. The Guardian tuvo acceso a un documento titulado “Intercept Partners”, donde se explica que algunas telecos habrían sido pagadas por su cooperación con la agencia.

[46] MacAskill, Ewen [èt.al]. *GCHQ taps fibre-optic cables for secret access to world’s communications* [en línea]. The Guardian, 2013 [Consulta: 6 junio 2016]. Disponible en: <https://www.theguardian.com/uk/2013/jun/21/gchq-cables-secret-world-communications-nsa>.

Sin embargo, el mismo artículo señala que una de sus fuentes del GCHQ afirma que las compañías en cuestión se vieron obligadas a cooperar forzosamente con la agencia y que se les prohíbe revelar la existencia de las órdenes que las obligan a dar acceso a sus cables.

Siguiendo con el artículo del The Guardian, podemos ver cómo se procesan los datos en TEMPORA.

Por un lado, sus sistemas de procesamiento aplican una serie de programas informáticos muy complejos con el fin de filtrar el material que entra. Este proceso se conoce como MVR, reducción masiva de volumen. Un primer filtro se encarga de reducir la entrada del tráfico que intercepta considerado como de bajo valor, como las descargas peer-to-peer[47], hasta reducir el volumen de entrada aproximadamente un 30%.

Otros filtros se encargan de generar paquetes de información con selectores fuertes: términos de búsqueda que incluyen temas, números de teléfono, direcciones de correos electrónicos e IPs. La mayor parte de la información resultante es contenido, como las grabaciones de llamadas telefónicas o los mensajes de los correos electrónicos, y el resto son metadatos.

[47] Red de ordenadores en la que todos o algunos aspectos funcionan sin clientes ni servidores fijos, sino mediante una serie de nodos que se comportan como iguales entre sí.

Bullrun y Edgehill

Dos PVMs pertenecientes a la NSA y a la GCHQ, respectivamente, para eludir el cifrado online y poder acceder a datos encriptados. Según el The Guardian[48], gracias a estos dos programas, tanto la NSA como la GCHQ habrían eludido gran parte de la codificación que salvaguarda sistemas bancarios y el comercio global, y que protege datos sensibles, como los secretos comerciales o los registros médicos.

Es un hecho muy grave para la privacidad de todos los internautas, ya que la confianza on-line reside precisamente en la criptografía.

No hay mucha información relativa a estos dos programas, sobre todo porque la mayoría de la documentación filtrada al respecto hace especial referencia a los acuerdos de las agencias con empresas tecnológicas para instalar debilidades en sus softwares de cifrado (puertas traseras). Sin embargo, tal y como se puede ver en una de las diapositivas filtradas pertenecientes a la GCHQ, gracias a Bullrun, sabemos que la NSA puede actuar contra los principales protocolos utilizados en línea, como HTTPS[49], voice-over-IP (VoIP)[50] o SSL[51].

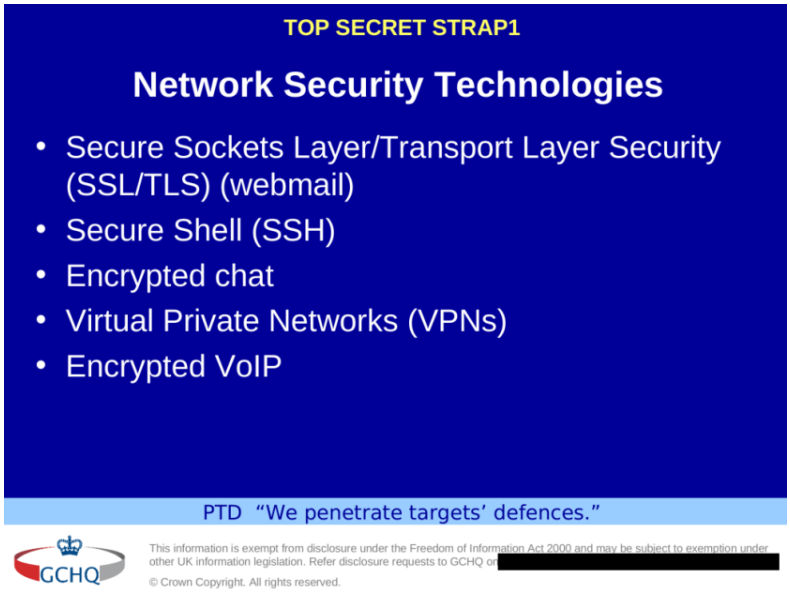

Tabla 33. Datos que pueden descifrar Bullrun y Edgehill (EdwardSnowden.com, 2015)

[48] Greenwald, Glenn; Ball, James; Borger, Julian. *Revealed: how US and UK spy agencies defeat Internet privacy and security* [en línea]. The Guardian, 2013 [Consulta: 21 junio 2016]. Disponible en: <https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security>.

[49] El protocolo de Transferencia de Híper-Texto (HTTPS) permite desarrollar actividades de e-commerce y cualquier transacción de forma segura.

[50] Conjunto de recursos que hacen posible que la señal de voz viaje a través de Internet empleando el protocolo IP (Protocolo de Internet).

[51] SSL, Secure Sockets Layer, es un protocolo diseñado para permitir que las aplicaciones transmitan información de ida y de manera segura hacia atrás.

Quantum Insert y FoxAcid

Quantum Insert es una técnica utilizada conjuntamente por la NSA y la GCHQ que permite romper la seguridad de múltiples sistemas para introducir malware[52] en ellos.

Quantum se basa en la redirección de servidores. Por ejemplo, cuando un usuario visita una página web, Quantum hace que el ordenador de dicho usuario cambie la ruta de la web original que estaba visitando y vuelque todo el tráfico en un servidor FoxAcid.

Un servidor FoxAcid puede introducir malware, monitorizar toda la actividad a tiempo real que esté llevando a cabo el objetivo infectado y copiar toda la información en una base de datos. Además, es capaz de redirigir el tráfico al servidor original de la web visitada, haciendo que el ataque sea muy complicado de detectar.

¿Cómo lo hacen? Los servidores FoxAcid poseen un software especial que permite elegir automáticamente el exploit a utilizar contra la víctima en cuestión. Un exploit es un fragmento de software, un fragmento de datos o una secuencia de comandos que se utiliza para aprovechar una vulnerabilidad de seguridad en un sistema de información para conseguir un comportamiento no deseado del mismo: acceso al sistema sin autorización, toma de control del sistema, acceso a privilegios no concedidos lícitamente por el sistema…

Para que el proceso sea automático parto de la idea de que existe una especie de catálogo de exploits. En base a ese catálogo y la importancia de la víctima, el propio servidor FoxAcid decide qué exploit es el más seguro de usar sin ser descubierto.

Por otro lado, los servidores FoxAcid disponen de otro software capaz de atacar un ordenador en forma de spam a través del navegador, desplegando una especie de "implantes" permanentes que infectan el equipo para dar acceso remoto completo a la máquina infectada.

[52] Malware es la abreviatura de "Malicious software", término que engloba a todo tipo de programa o código informático malicioso cuya función es dañar un sistema o causar un mal funcionamiento.

Utilizando este sistema, se sabe por las diapositivasfiltradas al respecto[53] que numerosos servicios son explotados por la NSA a través de Quantum: Alibaba, Hotmail, LinkedIn, Facebook, Twitter, Yahoo, Youtube, Microsoft o Gmail, entre otros.

Otro de los softwares que incluyen los servidores FoxAcid es el denominado Validator, un troyano que funciona a modo de puerta trasera y que se implanta a través de sistemas que operan con alguna versión de Windows.

Este troyano permite infectar un ordenador con malware más potente, como por ejemplo Olympusfire, un malware que permite tener el control total sobre un equipo: modificación, creación y copia de archivos, control sobre la webcam y los micrófonos, control de todas las comunicaciones que se realizan a través del equipo y copia de los nombres de usuario y contraseñas que utilice el usuario objetivo.

Según los documentos filtrados de la NSA sobre la red Tor[54], Quantum Insert es una herramienta muy útil para vulnerar la anonimidad de los usuarios de dicha red. La red Tor es una red de comunicaciones superpuesta a Internet donde se mantiene la integridad y el secreto de la información que viaja por ella y la anonimidad de la identidad IP de los usuarios.

Por otro lado, según el periódico The Wired[55], el programa se ha utilizado para hackear ordenadores de sospechosos de terrorismo, pero también contra los empleados de la compañía de telecomunicaciones Belgacom y los empleados de OPEC, la Organización de Países Exportadores de Petróleo.

---

[53] Spiegel Online. *NSA-Dokumente, so knackt der Geheimdienst Internetkonten* [en línea]. Der Spiegel, 2013. Diapositivas 13 y 14 [Consulta: 2 julio 2016]. Disponible en: <http://www.spiegel.de/fotostrecke/nsa-dokumente-so-knackt-der-geheimdienst-internetkonten-fotostrecke-105326-13.html>.

[54] The Guardian. *"Tor Stinks" presentation – read the full document* [en línea]. The Guardian, 2013 [Consulta: 2 julio 2016]. Disponible en: <http://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document>.

[55] Zetter, Kim. *How to detect sneaky NSA "Quantum Insert" attacks* [en línea]. The Wired, 2015 [Consulta: 2 julio 2016]. Disponible en: <https://www.wired.com/2015/04/researchers-uncover-method-detect-nsa-quantum-insert-hacks/>.

Dishfire

El lema de este PVM de la NSA lo dice todo: “mensajes de texto, una mina de oro por explotar”. Efectivamente, Dishfire es un PVM dedicado a recopilar y almacenar mensajes de texto. Los mensajes de texto que intercepta no siguen ningún criterio basado en personas sospechosas, sino que recopilan mensajes de texto de manera indiscriminada, tal y como informó el The Guardian[56] en enero de 2014. Solamente en el año 2011, Dishfire recopiló más de 194 millones de mensajes de texto al día.

Desde hace unos cuantos años los mensajes de texto han caído en picado como método de comunicación, sobre todo desde la aparición de los servicios de mensajería instantánea, como WhatsApp, un hecho que nos hace preguntar qué sentido tiene dedicar hoy en día un PVM única y exclusivamente a interceptar mensajes de texto. Sin embargo, las diapositivas nos muestran la verdadera razón:

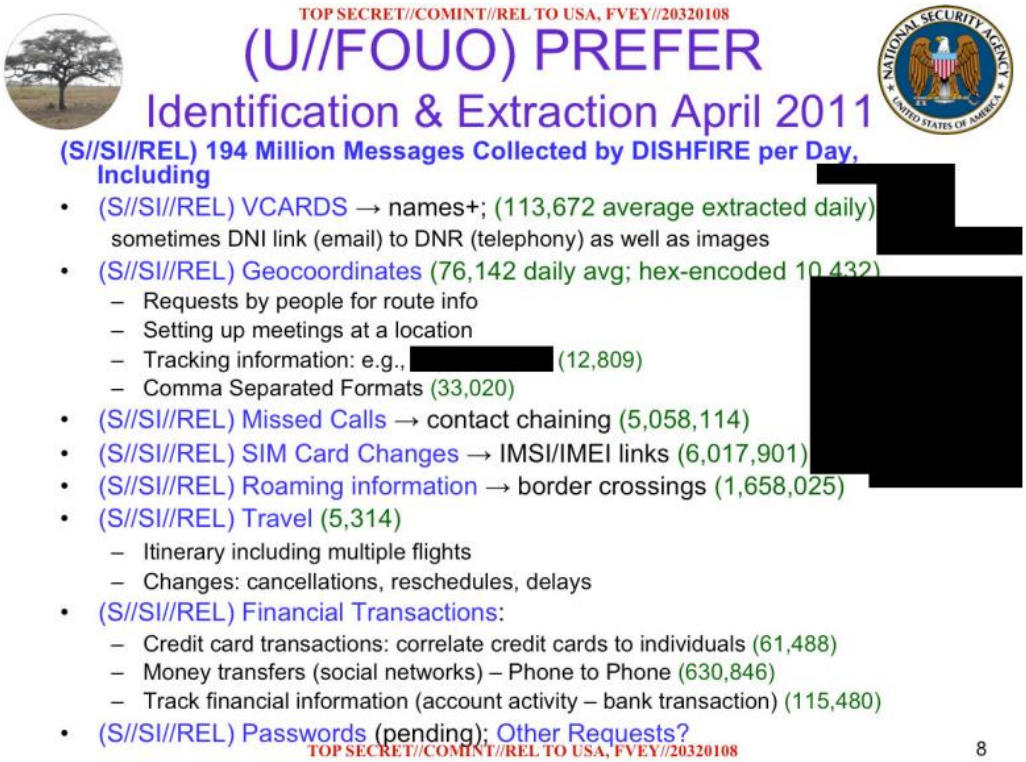

Tabla 34. Los datos que extrae Dishfire de los mensajes de texto. (EdwardSnowden.com, 2015)

[56] Ball, James. *NSA collects milions of text messages daily in “untargeted” global sweep* [en línea]. The Guardian, 2014 [Consulta: 4 julio 2016]. Disponible en: <https://www.theguardian.com/world/2014/jan/16/nsa-collects-millions-text-messages-daily-untargeted-global-sweep>.

Las comunicaciones entre personas es lo de menos. Lo que realmente interesa es información concreta: datos de avisos de llamadas perdidas para analizar la red de contactos de cada número, datos de transacciones financieras (pagos con tarjetas asociados a números de teléfono a los que se envían avisos), coordenadas geográficas e información automática que se envía a los teléfonos de los usuarios, como los cambios en el horario de un vuelo o cambios de contraseñas.

De esta manera, Dishfire es capaz de recopilar tanto los metadatos como el contenido de un mensaje de texto, tal y como podemos ver en la siguiente diapositiva:

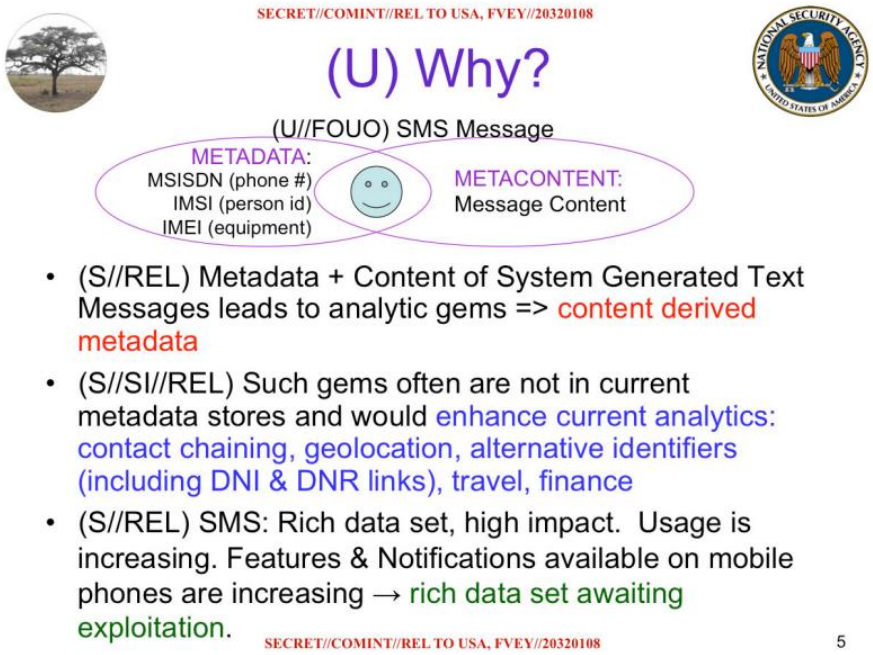

Tabla 35. Dishfire, recopilación de metadatos y contenido (EdwardSnowden.com, 2015)

En definitiva, Dishfire se encarga de recopilar automáticamente el contenido de los mensajes y tres tipos de metadatos concretos que ya hemos visto con anterioridad: el MSISDN, que es la identidad del abonado integrado en la red digital de servicios integrados; el IMSI, que corresponde a la identidad internacional del abonado a un móvil; y el IMEI, que es la identidad internacional del equipo móvil.

Además, a través de una serie de mejoras analíticas implantadas en Dishfire, otros datos que pasan a ser interceptados por este PVM son los relacionados con la cadena de contactos, la geolocalización, identificadores alternativos telefónicos y digitales, viajes y finanzas.

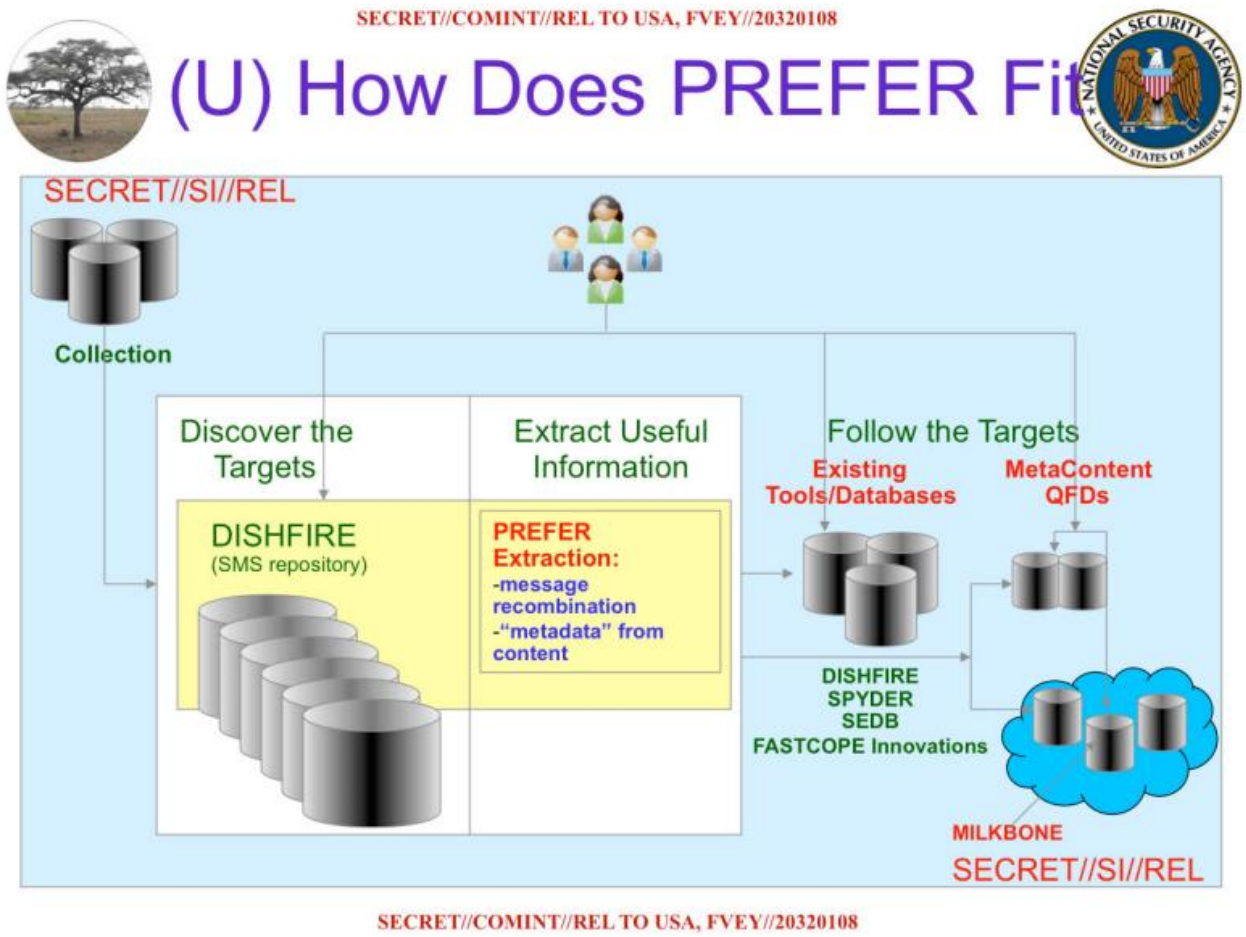

Tabla 36. Funcionamiento de Dishfire. (EdwardSnowden.com, 2015)

Dishfire identifica mensajes de texto e intercepta y extrae su contenido y metadatos, que pasan a almacenarse momentáneamente en un repositorio del propio programa. Los datos pasan a PREFER, una herramienta que actúa de filtro, para quedarse únicamente con la información que es de interés.

La información útil pasa entonces a almacenarse en diferentes bases de datos, según el tipo de dato que sea: en bases de datos ya existentes, en bases de datos de metadatos o en la base de datos Milkbone, especializada en almacenar colecciones de mensajes de texto.

Muscular

Se trata de un PVM usado conjuntamente por la NSA y la GCHQ, aunque su localización se encuentra en el Reino Unido y el responsable principal es la agencia inglesa. Muscular tiene asignado el Sigad[57] DS-200B, que significa que es de acceso internacional, pero de fuera de los Estados Unidos, lo que quiere decir que el programa opera a través de uno o varios cables de fibra óptica submarinos pertenecientes a alguno de los proveedores de telecomunicaciones de las agencias.

Básicamente, Muscular se encarga de interceptar el tráfico de datos que pasa por los cables de fibra óptica que transportan la información de los data centers de Yahoo! y Google. Interceptando estos cables, Muscular es capaz de acceder a los data centers de los dos gigantes informáticos y recolectar millones de registros al día, que incluyen tanto metadatos como contenido.

Es un PVM cuya existencia ha sorprendido a la opinión pública, ya que para interceptar los datos de Yahoo! y Google ya existe el programa PRISM, que es la joya de la corona. PRISM tiene acceso directo a los servidores de ambas empresas y es legal en los Estados Unidos, así que, ¿por qué crear otro programa que sea clandestino y tenga una función tan similar? Parece ser que, tal y como explica el Washington Post[58], Muscular es más agresivo que PRISM, precisamente porque funciona clandestinamente, sin necesidad de ninguna orden judicial, y porque el punto de acceso está fuera de Estados Unidos, de manera que no está bajo la jurisdicción del Tribunal de Vigilancia de Inteligencia Extranjera (FISC). Tiene menos restricciones y ninguna supervisión.

Cabe destacar que para evitar la pérdida de datos y ralentizaciones en sus sistemas, tanto Google como Yahoo! disponen de múltiples data centers repartidos por cuatro continentes y conectados entre ellos a través de cables de fibra óptica. Los datos fluyen sin problema entre los distintos data centers. Para que los data centers funcionen de manera efectiva, se sincronizan grandes volúmenes de información acerca de los titulares de las cuentas creadas.

---

[57] Sigad es la abreviatura de “signals intelligence address” o “signals intelligence activity designator”. Identifica la zona donde se recogen comunicaciones electrónicas.

[58] Gellman, Barton; Soltani, Ashkan. *NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say* [en línea]. The Washington Post, 2013 [Consulta: 15 julio 2016]. Disponible en: <https://www.washingtonpost.com/world/national-security/nsa-infiltrates-links-to-yahoo-google-data-centers-worldwide-snowden-documents-say/2013/10/30/e51d661e-4166-11e3-8b74-d89d714ca4dd_story.html>.

La red interna de Yahoo!, por ejemplo, transmite correos electrónicos enteros archivados durante años – años de mensajes y elementos adjuntos – desde un data center a otro.
Además, ambas compañías pagan por tener enlaces de datos de alta calidad, pensados para ser más rápidos, seguros y fiables, y durante los últimos años han comprado miles de quilómetros de cables de fibra óptica para su uso exclusivo. Sin embargo, de alguna manera, las agencias lograron eludir las potentes medidas de seguridad de los data centers de Yahoo! y Google. Tal y como se muestra en un boceto incluido en una de las diapositivas filtradas, hay un punto en que el "Internet público" se encuentra con "Google Cloud", la nube interna de Google donde residen sus datos. En el boceto se puede ver cómo se incluye una anotación indicando que el cifrado "¡se añade y se retira aquí!", junto a una cara sonriente.

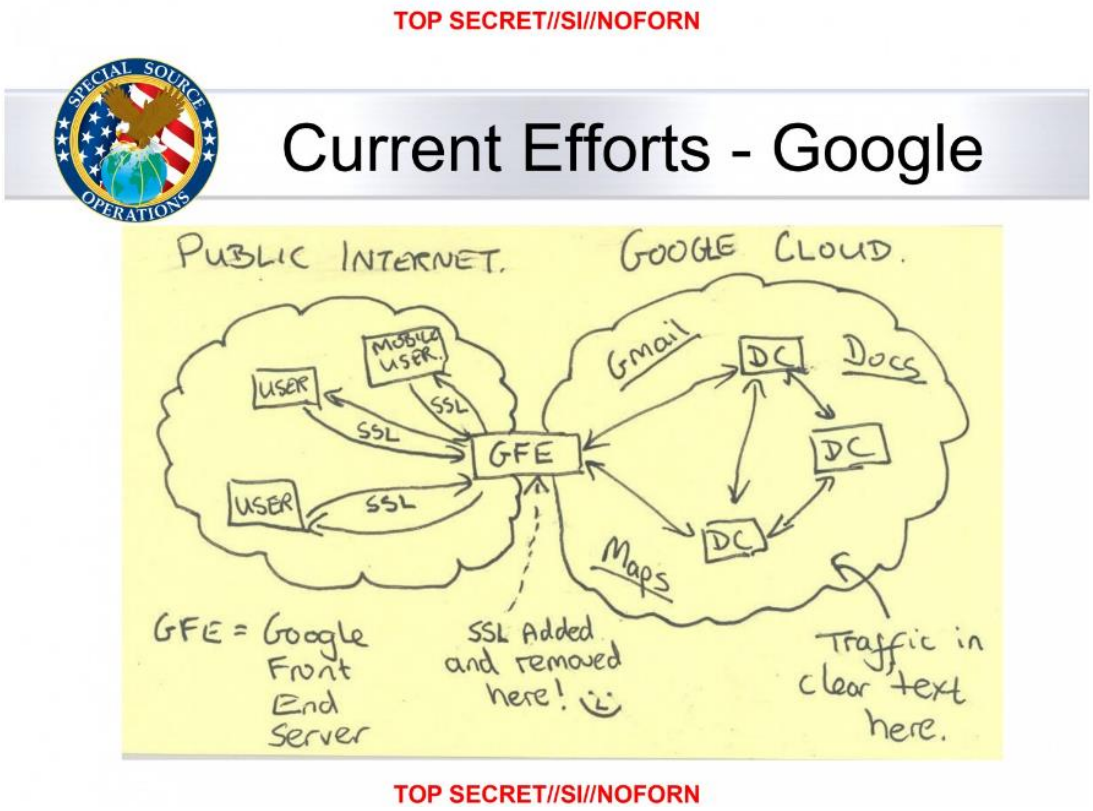

Tabla 37. Representación del punto de extracción de datos de Dishfire. (EdwardSnowden.com, 2015)

Una vez que Muscular tiene libre acceso a los data centers, recopila todo lo que encuentra y lo pasa a una especie de base de datos intermedia gestionada por la GCHQ, capaz de retener todo el tráfico realizado en un período de tres a cinco días. A continuación, la NSA utiliza una serie de herramientas echas a medida para decodificar los formatos especiales de los paquetes de datos que las dos compañías utilizan en sus nubes y los pasan por unos filtros, donde se selecciona la información que las agencias tiene marcada como útil y desechar la que no lo es.

Otros programas y actividades de interceptación de datos

Todos los PVM vistos con anterioridad tenían como protagonistas principales a la NSA y a la GCHQ. Evidentemente, son los casos más conocidos gracias a los documentos filtrados por Edward Snowden, donde la mayoría procedían precisamente de ambas agencias. Es por ello que, a pesar de no haber prácticamente información sobre los sistemas de interceptación de datos que utilizan las agencias de los Países Miembros, conviene no olvidar que estas dos agencias no son las únicas que invierten sus esfuerzos en el desarrollo de programas de vigilancia masiva.

Tenemos alguna información al respecto en un informe[59] realizado por Claude Moraes, de la Comisión de Libertades Civiles, Justicia y Asuntos de Interior del Parlamento Europeo. En él, destacan los casos de Suecia, Francia y Alemania.

Según el documento, gracias a los informes sobre las actividades realizadas por el Instituto de Radio Defensa Sueco (FRA), se ha descubierto que la agencia sueca recopila todos los datos que pasan por los cables de fibra óptica que atraviesan las fronteras del país procedentes de los países nórdicos, los estados bálticos y Rusia. Los informes también afirman que la agencia es capaz de controlar los datos y las llamadas de todos los móviles cuya señal se transmita a través de conexiones suecas.

En cuanto a Francia, se ha acusado a la Dirección General de Seguridad Exterior (DGSE) de interceptar y recopilar metadatos procedentes de correos electrónicos y mensajes de texto a través de estaciones satélite y de los cables submarinos de fibra óptica, y de enviar toda la información recabada a un superordenador.

Por otro lado, el servicio secreto alemán BND ha sido acusado de crear estaciones para desviar el tráfico entrante de Internet y poder copiarlo y analizarlo.

[59] Moraes, Claude. *Documento de Trabajo 1 sobre los programes de vigilancia de Estados Unidos y la UE y su repercusión sobre los derechos fundamentales europeos* [en línea]. Parlamento Europeo, 2015 [Consulta: 7 junio 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-524.799&format=PDF&language=ES&secondRef=01>.

Cesión

Gran parte de los datos recopilados por las agencias de inteligencia serían difícilmente recabados si no contaran con la inestimable ayuda de las grandes compañías de telecomunicaciones y de algunos gigantes de Sillicon Valley. Aunque tras las filtraciones de Snowden las grandes compañías de servicios de Internet negaron su participación en los programas de espionaje masivo (pese a que a finales de agosto de 2013 se publicó que habrían recibido millones de dólares[60] como contraprestación), la NSA volvió a insistir en marzo de 2014[61] de que sí son plenamente conscientes de ser parte del PVM PRISM.

De todas formas, aunque no se sabe a ciencia cierta si las grandes empresas de Internet señaladas en este trabajo consintieron voluntariamente la cesión de datos a la NSA a través de programas como PRISM, sí parece que hayan colaborado de un modo más silencioso a través de la implantación de puertas traseras en sus sistemas.

Tal y como reveló el The Guardian en septiembre de 2013[62], las agencias de inteligencia como la NSA o la GCHQ inglesa, en su afán por eludir el cifrado on-line, han adoptado una serie de métodos, además de uso de PVMs específicos como Bullrun o Edgehill. Estos métodos incluyen medidas encubiertas para asegurar su control sobre el establecimiento de normas internacionales de cifrado y la colaboración con empresas de tecnología y proveedores de servicios de Internet.

Esta colaboración permite a las agencias insertar vulnerabilidades en sus softwares de encriptación, o lo que se conoce como puertas traseras. Las puertas traseras permiten a las agencias acceder a información protegida que no podría estar disponible de otra manera.

---

[60] Timberg, Craig. *NSA paying US companyies for access to communications networks* [en línea]. The Washington Post, 2013 [Consulta: 2 julio 2016]. Disponible en: <https://www.washingtonpost.com/world/national-security/nsa-paying-us-companies-for-access-to-communications-networks/2013/08/29/5641a4b6-10c2-11e3-bdf6-e4fc677d94a1_story.html>.

[61] Ackerman, Spencer. *U.S. tech giants knew of NSA data collection, agency's top lawyer insists* [en línea]. Washington: The Guardian, 2014 [Consulta: 2 julio 2016]. Disponible en: <https://www.theguardian.com/world/2014/mar/19/us-tech-giants-knew-nsa-data-collection-rajesh-de>.

[62] Greenwald, Glenn; Ball, James; Borger, Julian. *Revealed: how US and UK spy agencies defeat internet privacy and security* [en línea]. The Guardian, 2013 [Consulta: 21 junio 2016]. Disponible en: <https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security>.

Las compañías mantienen que solamente colaboran con las agencias de inteligencia cuando están legalmente obligadas, sin embargo, según el The Guardian[63], Microsoft cooperó alegremente con la NSA debilitando la encriptación de los servicios de chat y e-mail de Outlook, aunque según la propia compañía, además de no existir puertas traseras en sus sistemas, las colaboraciones con agencias de inteligencias han sido fruto de obligaciones legales, y no por propia voluntad.

De todas formas, son muchas las diapositivas secretas de la NSA las que afirman que Microsoft colaboró encarecidamente con el FBI para desarrollar un sistema que permitiera acceder a las agencias a la información protegida de Outlook, Skype y SkyDrive.

Sea como sea, en ningún medio se tiene claro si las grandes compañías de Internet, exceptuando Microsoft, han colaborado con las agencias o han sido víctimas de ellas, aunque cabe destacar que en el Informe Moraes se las acusa de no contar con las medidas de protección adecuadas:

"…las empresas que han sido señaladas por los medios de comunicación por su participación en las operaciones de vigilancia masiva a gran escala de individuos europeos por parte de la Agencia Nacional de Seguridad de los EE.UU son empresas que han auto certificado su adhesión al principio de puerto seguro, y que el puerto seguro es el instrumento legal utilizado para la transferencia de datos personales de la Unión Europea a los Estados Unidos (por ejemplo, Google, Microsoft, Yahoo!, Facebook, Apple y LinkedIn); expresa su preocupación por el hecho de que estas organizaciones no hayan cifrado ni la información ni las comunicaciones que fluyen entre sus centros de datos, permitiendo de ese modo a los servicios de inteligencia interceptar información."

Más allá de las grandes compañías de servicios de Internet, son las grandes compañías de telecomunicaciones las que juegan un papel fundamental en este caso.

[63] Greenwald, Glenn [et.al]. *Microsoft handed the NSA access to encrypted messages* [en línea]. The Guardian, 2013 [Consulta: 28 junio 2016]. Disponible en: <https://www.theguardian.com/world/2013/jul/11/microsoft-nsa-collaboration-user-data>.

Algunos de los programas de los que se tiene constancia utilizados por la NSA para recopilar información a través de telecos son los pertenecientes al denominado grupo Upstream Collection, formado por los programas Blarney, Fairview, Oakstar y Stormbrew[64]. En una de las diapositivas sobre PRISM, se describe a Upstream como "una colección de comunicaciones procedentes de los cables de fibra óptica e infraestructura por donde fluyen datos".

A través de estos programas, la agencia saca provecho del acceso que ciertas telecos tienen a sistemas internacionales. La agencia dispone de un conjunto de contratos con las compañías mediante los cuales ellas desvían los datos de sus cables de fibra óptica a los data centers de la NSA.

Para la NSA, los programas de Upstream Collection y PRISM son sus dos principales medios de recogida de información y les dan a ambos la misma importancia, tal y como se puede ver en una de las diapositivas filtradas sobre PRISM:

Tabla 38. Upstream y Prism. (EdwardSnowden.com, 2015)

[64] Greenwald, Glenn. *Sin un lugar donde esconderse. Edward Snowden, la NSA y el Estado de vigilancia de EE.UU*. Nueva York: Metropolitan Books, 2014. Pág. 127. ISBN 978-84-666-5459-3.

Como podemos ver en la siguiente figura, los programas de vigilancia que integran Upstream Collection interceptan información procedente de los cables de fibra óptica, tanto de llamadas como de Internet. No almacenan los datos que interceptan, pero sí permiten el acceso a ellos a tiempo real.

Tabla 39. Características de los programas Upstream. (EdwardSnowden.com, 2015)

Más que la NSA, quien destaca sobre el resto de agencias en cuanto a asociaciones con empresas de telecomunicaciones se refiere, es la agencia de inteligencia inglesa GCHQ, tal y como reveló el The Guardian[65] en agosto de 2013. Según el artículo, las principales compañías de telecomunicaciones, como BT, Verizon y Vodafone, colaboran en secreto con la agencia, otorgándola acceso ilimitado a sus redes de cables submarinos.

En total se han identificado siete compañías de telecomunicaciones que colaboran con la agencia inglesa: BT, Verizon, Vodafone, Global Crossing, Level 3, Viatel e Interoute.

[65] Ball, James; Harding, Luke; Garside, Juliette. *BT and Vodafone among telecoms companyies passin details to GCHQ* [en línea]. The Guardian, 2013 [Consulta: 18 julio 2016]. Disponible en: <https://www.theguardian.com/business/2013/aug/02/telecoms-bt-vodafone-cables-gchq>.

En conjunto, estas siete empresas operan gran parte de los cables de fibra óptica submarinos que componen la columna vertebral de la arquitectura de Internet. Tal y como hemos visto en la explicación del PVM Tempora, en el 2010 la agencia tenía acceso a más de 200 cables y podía procesar al menos 46 a la vez.

Además, según el mismo artículo, el The Guardian tuvo acceso a una serie de documentos que revelaban que algunas de estas empresas de telecomunicaciones también dieron acceso a la agencia a cables que no son de su propiedad o que no operan ellos mismos.

En definitiva, gran parte de la información que nutre las bases de datos de las agencias de inteligencia proviene de asociaciones y contratos con empresas privadas de telecomunicaciones y con proveedores de servicios de Internet.

Tabla 40. Socios estratégicos de la NSA. (EdwardSnowden.com, 2015)

Compra

Además de la colaboración, ya sea obligada o no, con ciertas empresas de telecomunicaciones y de Internet, gran parte de los presupuestos de las agencias de inteligencia van a parar a contratistas de inteligencia y a empresas de ciberseguridad. Actualmente, la seguridad y los servicios de vigilancia asociados son un sector económico de grandes proporciones.

Según el libro de Glenn Greenwald "*Sin un lugar donde esconderse*", la NSA da empleo a unas 30.000 personas, pero tiene contrato con 60.000 más que pertenecen a compañías privadas. De hecho, según el analista Tim Shorrock, el setenta por ciento del presupuesto de inteligencia de la NSA se gasta en el sector privado. Asimismo, señala que en las inmediaciones de las oficinas de la NSA en Fort Meade, Maryland, se congregan una gran cantidad de contratistas vinculados con la Agencia, como Booz Allen Hamilton, SAIC o Northrop Grumman.

A finales de 2011, Wikileaks reveló una serie de documentación secreta y confidencial denominada Spyfiles[66] acerca de los contratistas de inteligencia que trabajan en la industria del espionaje masivo juntamente con agencias de inteligencia y gobiernos de todo el mundo. Son compañías procedentes de los países con la tecnología más sofisticada – especialmente Estados Unidos, Reino Unido, Alemania, Francia e Italia – que venden sus productos a los gobiernos y agencias de inteligencia que estén interesados, sin hacer distinciones.

Gracias a los documentos que recoge Spyfiles, un importante grupo de investigadores de la Universidad de Toronto denominado Citizen Lab, ha desarrollado diversos estudios al respecto, al igual que Reporteros sin Fronteras y su artículo sobre los enemigos de Internet[67]. De esta forma, podemos ver que dos grandes empresas privadas dedicadas al desarrollo de soluciones y tecnologías de vigilancia, como son Gamma International y Hacking Team, han vendido sus servicios a gobiernos como Estados Unidos, Alemania, Reino Unido, Italia y otros Estados Miembros.

---

[66] WikiLeaks. *The spy files* [en línea]. WikiLeaks, 2011 [Consulta: 15 julio 2016]. Disponible en: <https://wikileaks.org/the-spyfiles.html>.

[67] Reporters without Borders. *Enemies of the Internet. 2013 Report* [en línea]. París: International Secretariat Reporters without Borders, 2013 [Consulta: 16 julio 2016]. Disponible en: <http://surveillance.rsf.org/en/wp-content/uploads/sites/2/2013/03/enemies-of-the-internet_2013.pdf>.

Gamma International es una empresa con sede en Alemania especializada en vigilancia y monitoreo. Vende equipos y softwares destinados a vigilar y espiar, como por ejemplo FinFisher, un malware que infecta ordenadores para poder tener acceso remoto a toda su actividad en tiempo real.

Gamma International fue incluido en los Spyfiles en el año 2011 y su software FinFisher fue denunciado por Mozilla al descubrir que se hacía pasar por un complemento de su navegador Firefox. Según un estudio realizado por Citizen Lab en abril de 2013, "*For Their Eyes Only: The Commercialization of Digital Spying*"[68], FinFisher tiene presencia en más de 35 países, un hecho que sugiere la posibilidad de que el software haya sido adquirido por los gobiernos o agencias de esos países. Destacan Estados Unidos y un gran número de Países Miembros, como Austria, Alemania, Bulgaria, República Checa, Estonia, Lituania, Letonia, Hungría, Holanda, Rumanía y el Reino Unido.

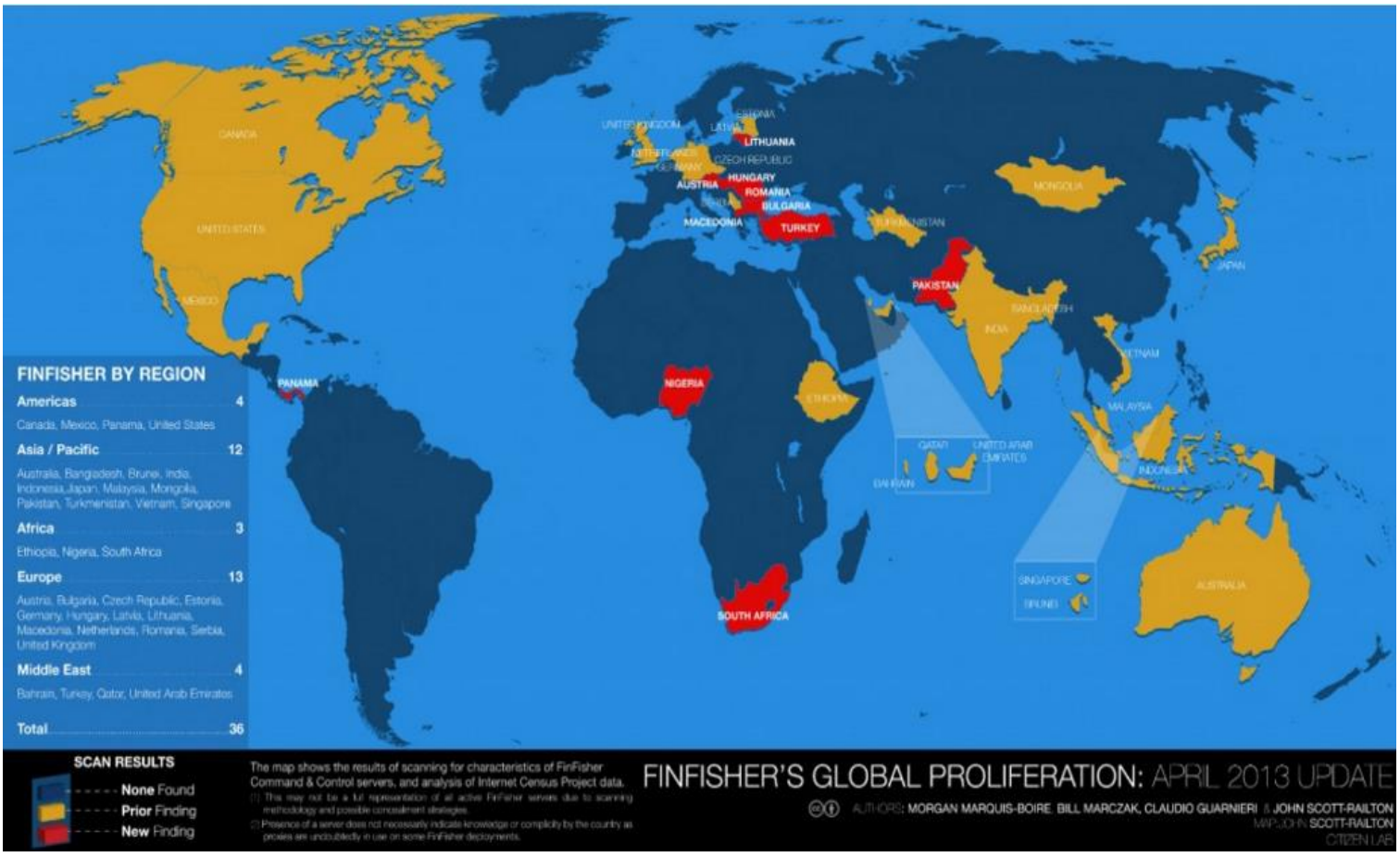

Tabla 41. Proliferación del programa FinFisher. (The Citizen Lab, 2013)

[68] The Citizen Lab. *For Their Eyes Only: The Commercialization of Digital Spying* [en línea]. Toronto: Munk School of Global Affairs, University of Toronto, 2013 [Consulta: 15 julio 2016]. Disponible en: <https://citizenlab.org/2013/04/for-their-eyes-only-2/>.

Por su parte, Hacking Team es una compañía italiana que vende herramientas de vigilancia e intrusión a gobiernos y agencias de inteligencia. Disponen de una amplia gama de sistemas de control remotos que permiten controlar cualquier comunicación de un usuario en Internet, descifrar criptografías de archivos y correos electrónicos, grabar llamadas de Skype y de otras comunicaciones de VoIP, extraer contraseñas y activar remotamente micrófonos y cámaras de ordenadores.

Citizen Lab calcula que los productos de Hacking Team tienen presencia en más de 20 países, entre ellos algunos Países Miembros, como Italia, Hungría y Polonia.

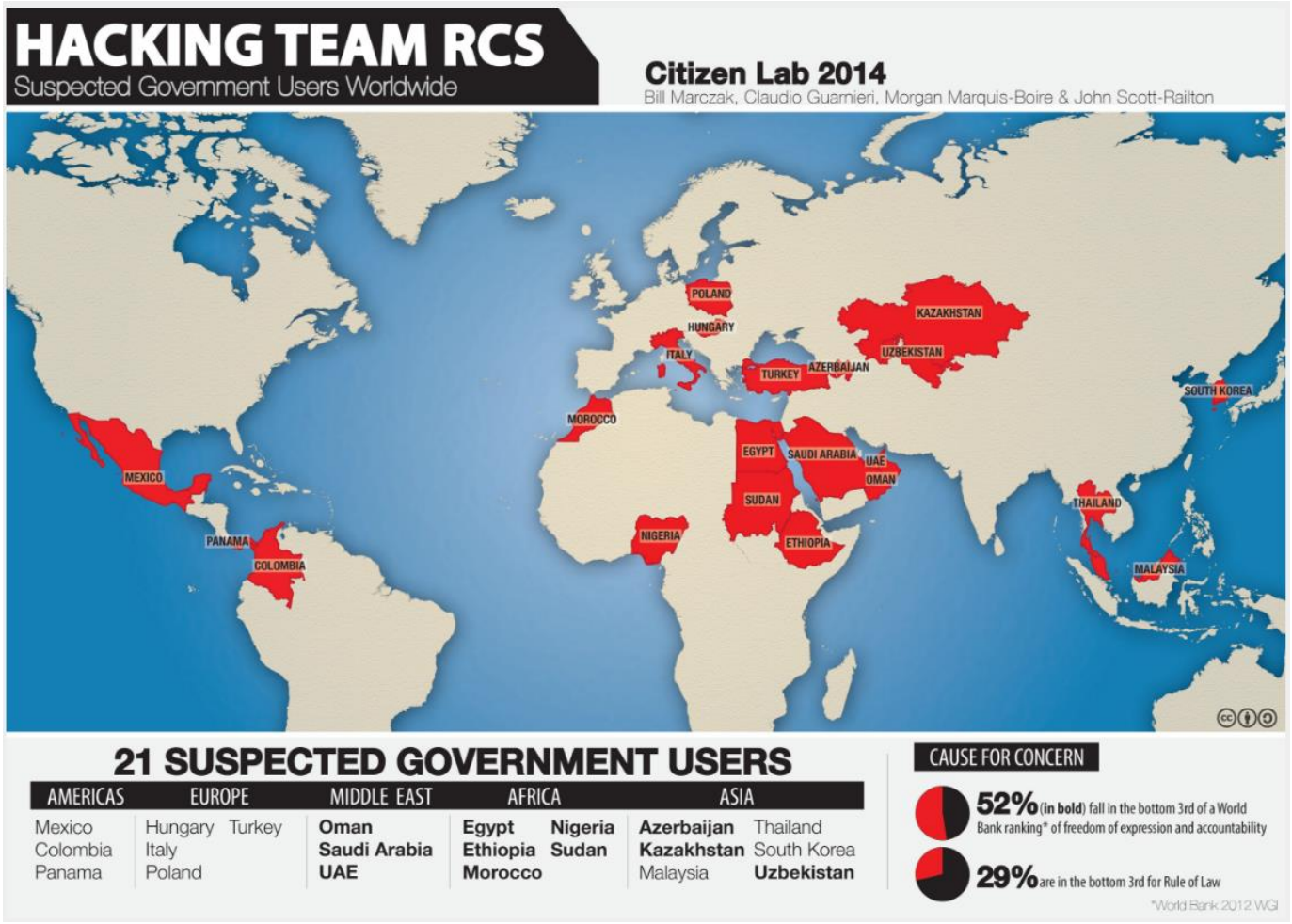

Tabla 42. Gobiernos sospechosos de utilizar productos de Hacking Team (The Citizen Lab, 2014)

Colaboración entre agencias

Cuando saltaron a la luz pública los documentos secretos de la NSA revelados por Edward Snowden, la prensa se hizo eco de una serie de diapositivas sobre un programa de análisis de la NSA llamado Boundless Informant. En estas diapositivas aparecían diversos gráficos que indicaban el número de metadatos telefónicos que la NSA había recopilado entre diciembre de 2012 y enero de 2013 de múltiples Países Miembros, como Francia, España, Italia, Holanda o Alemania.

En un principio, la interpretación de los medios de comunicación fue la de pensar que la NSA había recopilado esta información a través de alguno de sus múltiples PVM. Sin embargo, meses después de las filtraciones, el director de la NSA, Keith Alexander, afirmaba que precisamente esos metadatos no habían sido recopilados por la NSA o la agencia inglesa, sino que las encargadas de dicha tarea fueron las propias agencias de inteligencia de cada uno de los países que aparecían en las diapositivas y que posteriormente compartieron con la NSA[69].

De todas formas, hay muchas evidencias que confirman la colaboración entre distintas agencias europeas con la NSA, sobre todo en lo que a compartir datos de ciudadanos se refiere. Sin tener en cuenta a Gran Bretaña, que como ya hemos visto forma parte de los Cinco Ojos y está presente en prácticamente todos los PVM de la NSA, las colaboraciones que más destacan son las de las agencias alemana, francesa y española.

En el informe Moraes también queda patente esta colaboración, pues recordemos que en el punto 22 del apartado de Principales Conclusiones se pide a los Estados Miembros que revisen las prácticas que sus agencias de inteligencia llevan a cabo junto con las agencias norteamericanas.

Si comenzamos por Alemania, podemos ver que, según los principales medios de comunicación alemanes, la NSA y la agencia alemana BND mantienen una relación mucho más estrecha de lo que se pensaba. Para empezar, la BND transfiere sistemáticamente metadatos recopilados por ellos mismos a la NSA.

---

[69] Caño, Antonio. *La NSA afirma que el espionaje masivo fue realizado por Francia y España* [en línea]. El País, 2015 [Consulta: 4 julio 2016]. Disponible en: <http://internacional.elpais.com/internacional/2013/10/29/actualidad/1383074305_352044.html>.

Según el periódico alemán Der Spiegel[70], durante el mes de diciembre de 2012 pasaron alrededor de 500 millones de registros de metadatos a la NSA. A cambio, la NSA provee acceso a la agencia alemana al PVM XKeyScore, y a su vez, la BND da acceso a la NSA a los PVM alemanes Mira4 y VERAS[71], de los que por desgracia no se tiene información.

Esta colaboración entre ambas agencias se produce regularmente desde hace décadas[72], al igual que sucede con cinco Países Miembro más: España, Italia, Francia, Reino Unido y Holanda.

En cuanto a España, tal y como reveló El Mundo[73] en octubre de 2013, la intrusión de la NSA en la vida cotidiana de millones de españoles fue más fruto de la cooperación entre ambos países que de un abuso de poder norteamericano. Los servicios de inteligencia españoles no solamente conocían el trabajo de la NSA, sino que además le facilitaban sus tareas. El CNI habría ayudado a intervenir a la NSA 60 millones de llamas telefónicas entre diciembre de 2012 y enero de 2013.

Cabe destacar además un artículo de la periodista Magda Bandera para la revista Playboy en 2003[74] donde se afirma que el expresidente norteamericano George Bush ofreció en 2001 a José María Aznar, presidente de España en aquella época, compartir la red de inteligencia Echelon, tal y como confirmó también el The Guardian[75] dos años antes. Echelon era considerada la mayor red de espionaje de interceptación de comunicaciones electrónicas de la historia, hasta las revelaciones de Snowden. El famoso PVM PRISM forma parte de la trama de vigilancia de Echelon, por lo que se considera altamente probable que España tenga o haya tenido acceso al PVM PRISM.

---

[70] Spiegel Online. *Überwachung: BND leitet massenhaft Metadaten an die NSA weiter* [en línea]. Hamburg: Der Spiegel, 2013 [Consulta: 4 julio 2016]. Disponible en: <http://www.spiegel.de/netzwelt/netzpolitik/bnd-leitet-laut-spiegel-massenhaft-metadaten-an-die-nsa-weiter-a-914682.html>.

[71] Poitras, Laura; Gude, Hubert; Rosenbach, Marcel. *Mass Data: transfers from Germany Aid US Surveillance* [en línea]. Spiegel Online International, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://www.spiegel.de/international/world/german-intelligence-sends-massive-amounts-of-data-to-the-nsa-a-914821.html>.

[72] The Copenhagen Post Online. *Denmark in US spy agreement?* [en línea]. The Copenhagen Post, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://cphpost.dk/news/international/denmark-in-us-spy-agreement.html>.

[73] Greenwald, Glenn; Aranda, Germán. *El CNI facilitó el espionaje masivo de EEUU a España* [en línea]. El Mundo, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://www.elmundo.es/espana/2013/10/30/5270985d63fd3d7d778b4576.html>.

[74] Bandera, Magda. Lo que el sistema sabe sobre ti. *Revista Playboy*. 2003. Núm. 3, época 2.

[75] Tremlett, Giles. *US offers to spy on ETA for Spain* [en línea]. The Guardian, 2001 [Consulta: 9 julio 2016]. Disponible en: <https://www.theguardian.com/world/2001/jun/15/spain.usa>.

Respecto a Francia, destaca la información proporcionada por el diario francés Le Monde[76], donde se explica que durante el 2011, la agencia francesa DGSE y la NSA firmaron un memorándum de intercambio de datos, facilitando de esta manera la transferencia de millones de registros de metadatos de ciudadanos franceses. En un mes, más de 70 millones de registros de metadatos fueron facilitados a la NSA. Este acuerdo de cooperación firmado entre la agencia francesa y los miembros de los Cinco Ojos recibe el nombre de Lustre.

El acuerdo secreto denominado Lustre surgió durante el 2006. El interés de la NSA en este acercamiento recae principalmente en la gran posición geoestratégica que tiene Francia respecto al tráfico de datos electrónicos, ya que las comunicaciones que unen África y Europa pasan por cables submarinos que entran en las costas de Penmarch y Marsella. La DGSE intercepta los datos que se transmiten por esos cables, y según el acuerdo Lustre, se los pasa tanto a la NSA como a las otras cuatro agencias que forman parte de los Cinco Ojos.

4.4 Procesamiento, almacenamiento y análisis de datos

Una vez recopilada la información a través de los distintos medios que hemos podido ver en el apartado anterior, ¿qué proceso siguen los datos hasta acabar siendo analizados en una base de datos?

Para explicarlo me fijaré únicamente en los procesos documentados de la NSA, ya que del resto de agencias de inteligencia no hay ningún tipo de información disponible. A pesar de que la NSA es la agencia de inteligencia más potente del mundo, cuyo presupuesto sobrepasa exageradamente al que reciben el resto de agencias aquí estudiadas, el proceso relativo al procesamiento, el almacenamiento y el análisis de los datos sirve para poder hacernos una idea general de cómo realizan todo esto el resto de agencias.

[76] Follorou, Jacques. *Surveillance: la DGSE a transmis des données à la NSA américaine* [en línea]. Le Monde, 2013 [Consulta: 9 julio 2016]. Disponible en: <http://www.lemonde.fr/international/article/2013/10/30/surveillance-la-dgse-a-transmis-des-donnees-a-la-nsa-americaine_3505266_3210.html>.

4.4.1 De la extracción al almacenamiento

El proceso desde que se extraen los datos hasta que se almacenan en una base de datos definitiva es, en esencia, el siguiente:

1. Automatización del tráfico de datos
2. Filtrado y clasificación de los datos
3. [Sólo metadatos] Procesamiento de los metadatos
4. [Sólo metadatos] Base de datos intermedia de encadenamiento de metadatos
5. Almacenamiento en la base de datos definitiva

Si la recolección la realiza un tercero, el proceso es prácticamente el mismo, con la diferencia de que antes de llegar al filtrado y a la clasificación de los datos es necesario encriptar y transferir la información a la agencia.

1. Automatización del tráfico de datos / Transferencia encriptada

Si la recolección de los datos la realiza la propia agencia, tenemos como ejemplo en este primer paso al sistema Printaura, que se encarga de automatizar todo el tráfico de datos que proviene del PVM PRISM. Concretamente, Printaura distribuye el flujo de datos en función de si son datos de voz, texto, vídeo o metadatos y asigna las tareas específicas que debe seguir el sistema a lo largo de todo el proceso.

Si la recolección la realiza un tercero, el primer paso es la transferencia de los datos a la agencia. En este caso tenemos al programa Mailorder, un sistema que transfiere datos de forma encriptada.

2. Filtrado y clasificación

Aquí destacan los programas Courierskill y Scissors.
Courierskill es un filtro que se encarga de seleccionar aquellos datos de tipo contenido que son de interés para su posterior análisis. Los datos que determina relevantes pasan a la siguiente fase, desechando los que no han sido seleccionados.

Por su parte, a Scissors lo encontramos en el proceso que sigue la información que proviene de PRISM. Tras el paso de los datos por Printaura, éstos van a parar a Scissors, que se encarga de darles un formato inicial y clasificarlos según sus características para determinar en qué base de datos se tienen que almacenar.

3. Procesamiento de los metadatos

Los datos recolectados que sean metadatos pasan por otros dos programas, dependiendo de su procedencia: Fallout y Fascia.

Fallout es un sistema que la NSA define como "ingest processor". No estoy segura del significado del término, pero en mi búsqueda he encontrado un plugin de Apache[77] llamado *ingest attachment processor* que parece muy similar. Si tenemos en cuenta el plugin de Apache, podríamos aventurarnos a definir a Fallout como un sistema convertidor que proporciona a los metadatos recopilados un formato común y fácilmente legible, como pdf o xls. De esta forma, los metadatos se pueden volcar en una base de datos definitiva con unos formatos legibles, tanto para la máquina como para el analista.

Fascia es exactamente el mismo tipo de programa, con la diferencia de que mientras Fallout recibe solamente los metadatos procedentes de Internet, Fascia recibe los que proceden de llamadas y de mensajes de texto.

4. Base de datos intermedia

Tras dar a los metadatos unos formatos legibles, éstos pasan a una base de datos intermedia. Como ejemplo tenemos a la base de datos Mainway, definida por la NSA como "chaining database".

En Mainway se almacenan metadatos telefónicos y metadatos de correos electrónicos, y se encarga de establecer una relación entre ambos tipos de metadatos. Por ejemplo, puede relacionar un número de móvil con una cuenta de correo electrónico.

Es decir, permite a un analista identificar cadenas de comunicación que fluyen por distintas redes de telecomunicaciones. Personalmente, opino que es un sistema que, en resumidas cuentas, permite contextualizar datos que por sí solos no tienen ningún sentido. Lo curioso es que parece ser que solamente almacena estos vínculos, porque los metadatos en sí se conservan en bases de datos definitivas.

5. Almacenamiento en una base de datos definitiva

La NSA tiene casi un centenar de bases de datos dedicadas a almacenar y conservar todos los datos recopilados en función de su tipología, características o procedencia.

[77] Elastic. *Ingest attachment processor Plugin* [en línea]. Elasticsearch, 2016 [Consulta: 15 julio 2016]. Disponible en: <https://www.elastic.co/guide/en/elasticsearch/plugins/master/ingest-attachment.html>.

A diferencia de las bases de datos utilizadas por las grandes compañías de Internet que hemos podido ver en apartados anteriores, en este caso no se tiene apenas información sobre ellas, de manera que su capacidad de almacenamiento y su período de conservación de los datos son, en la mayoría de los casos, desconocidos.

De todas formas, algunas de las bases de datos más destacadas gracias a su aparición en los medios de comunicación a raíz de las filtraciones de Snowden son las siguientes:

Base de datos	Contenido /Metadatos	Procedencia datos	Tipo de datos	Periodo de conservación
Nucleon	Contenido	DNR* y DNI*	Datos de voz	30 días
Association	Metadatos	DNR	Llamadas telefónicas de móviles	Desconocido
Banyan	Metadatos	DNR	Llamadas telefónicas de teléfonos fijos	Desconocido
Marina	Metadatos	DNI	Datos de navegación, contactos, comunicaciones	1 año
Traffichief	Metadatos	DNR y DNI	Selectores fuertes (direcciones e-mail, núm. teléfonos, IPs)	Desconocido
Pinwale	Contenido	DNI	Chats, e-mails, vídeos, foros, etc	5 años
Dishfire	Contenido y metadatos	DNR	Mensajes de texto	Desconocido
XKeyScore / Tempora	Contenido y metadatos	DNR y DNI	Los recolectados por ellos mismos	3 días contenido, 30 metadatos
Tracfin	Contenido y metadatos	DNI	Transacciones con tarjetas de crédito	Desconocido

DNR: Dialed Number Recognition, datos telefónicos

DNI: Digital Network Intelligence, cualquier actividad realizada vía Internet.

A pesar de no disponer de ningún tipo de información técnica, cabe suponer que las distintas bases de datos de la NSA son de tipo NoSQL si tenemos en cuenta la enorme cantidad de datos desestructurados que procesan, a excepción tal vez de bases de datos como Banyan o Association, que recogen metadatos concretos procedentes del mismo tipo de fuentes.

Otra cuestión destacable es que, probablemente, la gran mayoría de bases de datos tengan la capacidad de interactuar entre ellas para garantizar una recuperación completa de resultados.

Una prueba de ello la encontramos en un documento que sirve de manual de usuario acerca de los datos recopilados de Skype a través de PRISM[78], que nos muestra una interacción entre las bases de datos Pinwale y Nucleon:

f. Where's the audio to go with my Skype video?

f.i. Within the DNI Presenter (DNIP), the user can utilize the "View Associations" service to find the associated NUCLEON audio of the PINWALE document or find the associated PINWALE document of the NUCLEON audio. In other words, you can find the MAM and/or TAM associations to any associated Skype data, and then display them within the DNIP Skype combined display or the DNIP Composite display. Also from the UIS Text Presenter, you can launch the DNI Presenter to utilize this "View Associations" service.

Tabla 43. ¿Dónde está el audio relacionado con mi vídeo de Skype? (EdwardSnowden.com, 2015)

Esto quiere decir que Pinwale cuenta con una herramienta que permite ver asociaciones, en este caso contenido relacionado alojado en Nucleon, y viceversa. De esta forma, un analista puede ver un chat de Skype en Pinwale y escuchar el audio de la llamada correspondiente en Nucleon.

[78] Snowden Doc Search. *User's Guide for PRISM Skype Collection* [en línea]. Journalistic Source Protection Defence Fund, 2012 [Consulta: 20 julio 2016]. Disponible en: <https://search.edwardsnowden.com/docs/User%E2%80%99sGuideforPRISMSkypeCollection2014-12-28nsadocs>.

Ejemplos de procesos

Para poder ejemplificar todos los pasos anteriormente explicados que se llevan a cabo durante el procesamiento y el almacenamiento de la información recopilada, nos fijaremos en el recorrido de los metadatos en general que recolectan los socios de la NSA, en los metadatos DNI que se recogen a través de los programas de Upstream y en el flujo de datos de PRISM.

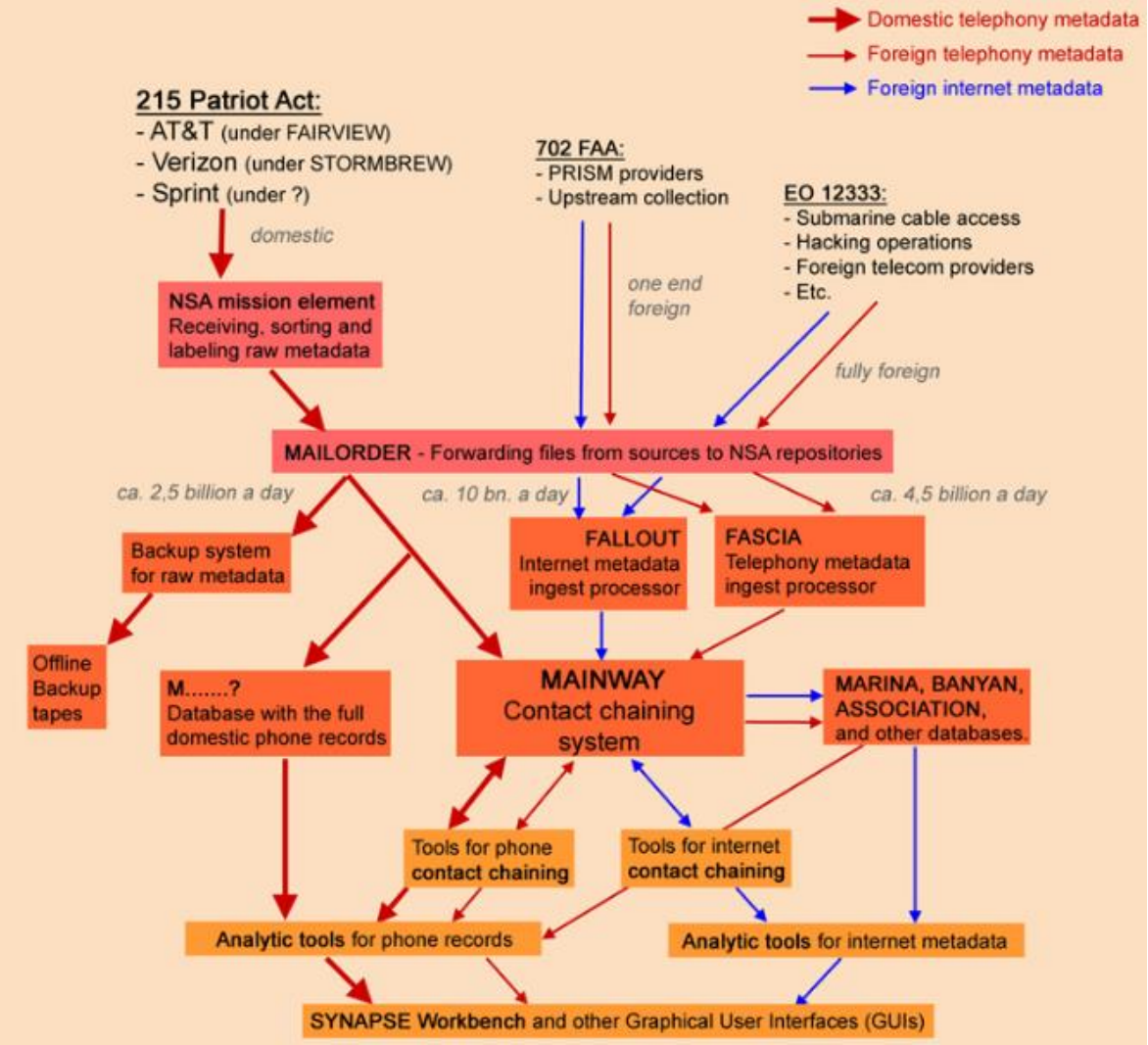

Tabla 43. Flujo de los metadatos recolectados por terceros. (Electrospaces, 2014).

En esta tabla se muestra el flujo general que siguen los metadatos que son recolectados por los socios y proveedores de la NSA. El proceso es el siguiente:

1. Las compañías de telecomunicaciones asociadas y los proveedores de PRISM transfieren los metadatos recolectados a Mailorder, el sistema encargado de transferir la información de forma encriptada a los repositorios de la NSA.

2. Una vez en manos de la NSA, los metadatos procedentes de Internet son absorbidos por Fallout, y los procedentes de sistemas de telefonía por Fascia, que se encargan de proporcionarles formatos legibles.

3. Los metadatos se almacenan temporalmente en Mainway para poder establecer relaciones entre ellos antes de que pasen a ser almacenados en diferentes bases de datos.

4. Finalmente, en función de su tipología, los metadatos pasaran a almacenarse en bases de datos definitivas. Por ejemplo, si son metadatos procedentes de llamadas telefónicas realizadas con móviles se almacenan en Association, o si son metadatos procedentes de Internet se almacenan en Marina.

En la siguiente tabla se muestra el flujo de los metadatos procedentes de Internet (DNI) que son recolectados por los programas de Upstream, es decir, por socios de telecomunicaciones:

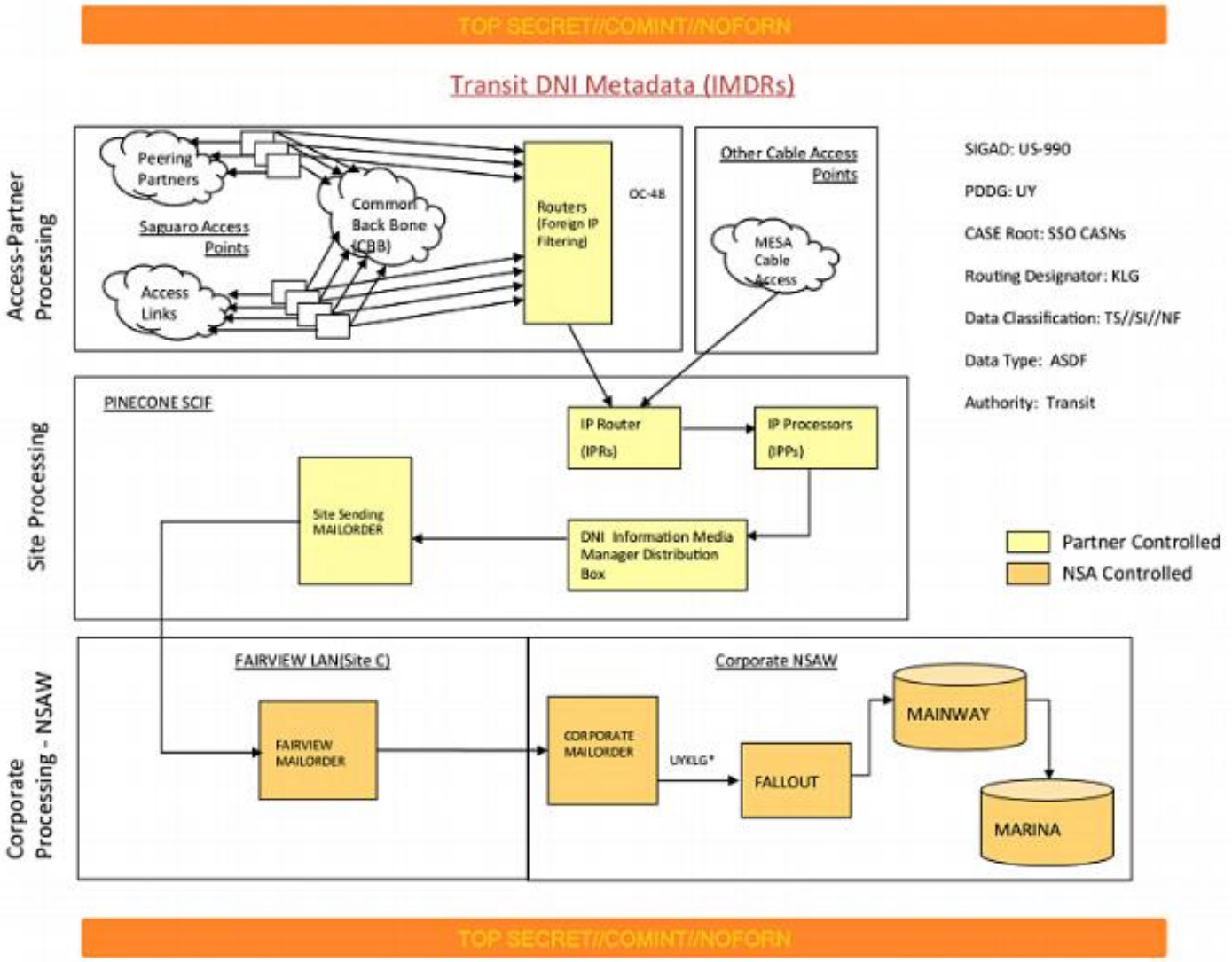

Tabla 44. Flujo de los metadatos DNI recolectados por los programas de Upstream. (EdwardSnowden.com, 2015)

En primer lugar, el diagrama nos muestra que el proceso se lleva a cabo en tres localizaciones diferentes:

1. Compañía socia
2. Sitio de procesamiento seguro
3. Repositorios de la NSA

En cada uno de estos sitios ocurre lo siguiente:

1. Procesamiento en la compañía socia

- La compañía extrae los metadatos DNI a partir de varios accesos a sus cables de fibra óptica, entre los que se incluyen aquellos de su propiedad que usan específicamente otras empresas de telecomunicaciones. Es lo que se conoce como Back Bone, es decir, el eje principal de los datos que fluyen por los cables.
- Todos los metadatos extraídos de los diversos puntos convergen en uno o varios rúters encargados de seleccionar los que hagan referencia a ciudadanos no estadounidenses.

2. Sitio de procesamiento seguro

- Los metadatos DNI recopilados por la compañía pasan a una central de alta seguridad gestionada tanto por personal de la teleco como de la NSA.
- A través de algo denominado como rúters IP y procesadores IP, los metadatos están listos para ser transferidos a Mailorder.
- En Mailorder los metadatos son encriptados para poder ser transferidos a los repositorios de la NSA con seguridad. Si en lugar de metadatos hablásemos de contenido, la información tendría que pasar antes por Courierskill, el sistema que actúa a modo de filtro.

3. Procesamiento y almacenamiento en los repositorios de la NSA

- Los metadatos llegan a la red local de la NSA a través de Mailorder.
- De la red local pasan a la red corporativa, utilizando de nuevo Mailorder.
- Los metadatos son procesados por Fallout para darles un formato legible.
- Tras Fallout, se almacenan temporalmente en Mainway, donde se establecen las múltiples y diferentes relaciones existentes entre ellos.
- Finalmente, al tratarse de metadatos procedentes de las actividades en Internet, se almacenan en la base de datos Marina.

Como último ejemplo tenemos el flujo de datos en el PVM PRISM, muy parecido a los anteriores, con la novedad de que en esta ocasión hay datos de tipo contenido.

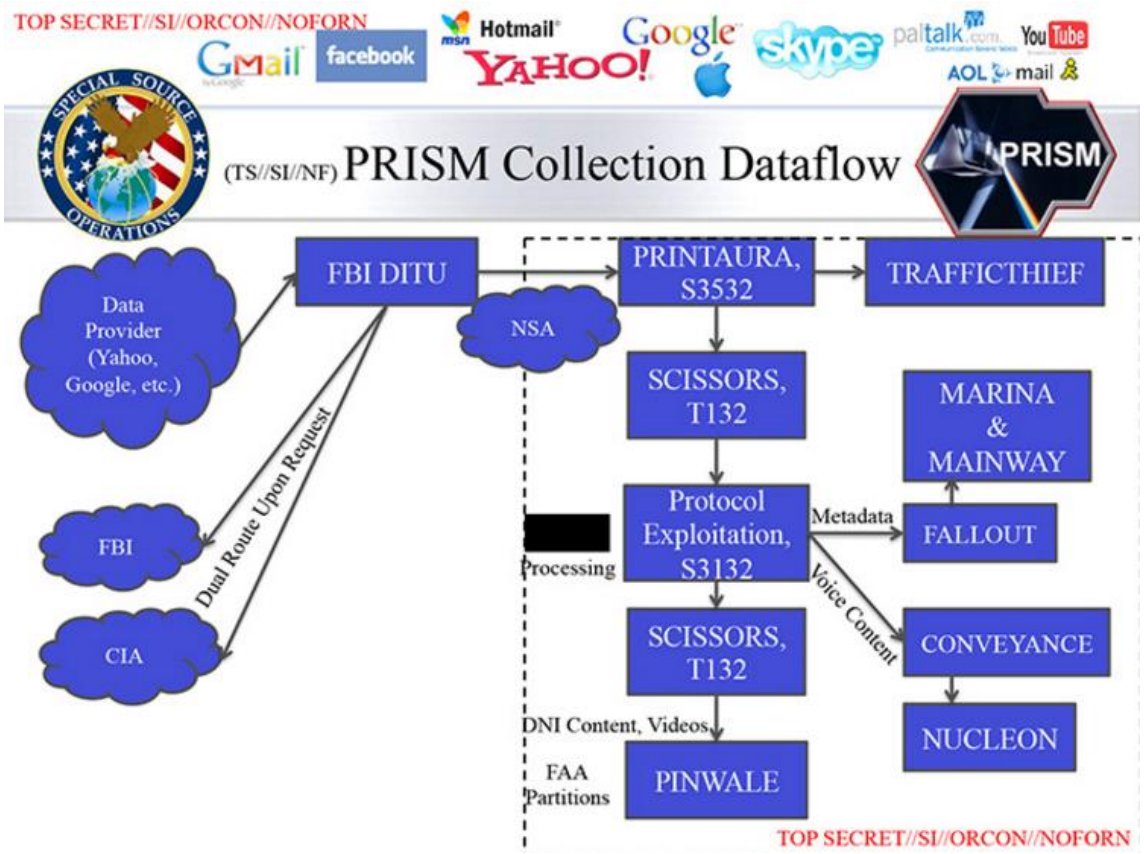

Tabla 45. Flujo de datos en PRISM. (EdwardSnowden.com, 2015)

1. La Unidad Tecnológica de Interceptación de Datos (DITU) del FBI es la encargada de recolectar los datos procedentes de los proveedores de PRISM (Google, Yahoo, Microsoft, Facebook, etc.).

2. El DITU transfiere a la NSA los datos interceptados, que pasan en primera instancia por el sistema Printaura para automatizar el tráfico de datos, es decir, para distribuir y asignar tareas automáticamente.

3. Todos los datos denominados como selectores fuertes (números de teléfono, direcciones IP, direcciones de correo electrónico, nombres de usuarios…) detectados por Printaura pasan automáticamente a la base de datos Traffichief. El resto van a parar al sistema Scissors.

4. Una vez en Scissors, los datos son clasificados para saber en qué base de datos deben almacenarse.

5. Tras su clasificación en Scissors, los datos pasan por una unidad denominada Protocol Explotation, de la que no hay ningún tipo de información al respecto, así que no se sabe qué es lo que se hace con los datos en esta fase, pero parece que desde ahí son redireccionados a diferentes sistemas, dependiendo de si son metadatos, contenido o datos de voz.

6. Si son metadatos se procesan a través de Fallout, que les da un formato legible y los manda a Mainway, que establece relaciones entre los distintos metadatos antes de enviarlos a la base de datos definitiva Marina.

7. Si son datos de voz, éstos se procesan a través de Conveyance, el equivalente a Fallout, que tras darles un formato los almacena en la base de datos especializada en datos de voz Nucleon.

8. En caso de que la información sea de tipo contenido, ésta vuelve a pasar por el sistema Scissors, supongo que para precisar aún más su clasificación, y de ahí pasa a almacenarse en bases de datos definitivas como Pinwale.

4.4.2 Análisis de los datos

A la hora de analizar los datos, la NSA cuenta con diversos sistemas integrados en sus bases de datos que permiten buscar, cotejar, monitorear y cruzar distintos tipos de datos alojados en las distintas bases de datos. También cuentan con programas que analizan conjuntos de datos muy específicos de una manera visual y sencilla, como es el caso de los programas Boundless Informant y Unified Targeting Tool (UTT). Así mismo, se conoce la existencia de un sistema denominado Elegantchaos, que se menciona de pasada en algunas diapositivas sobre PRISM, y que supuestamente es un sistema de análisis de datos a gran escala, aunque por desgracia no se sabe nada más al respecto. Sin embargo, la pieza clave en el análisis de los datos parece ser un programa denominado Accumulo, una gigantesca base de datos similar a la BigTable de Google.

Boundless Informant

Se trata de un software analítico basado en las tecnologías Big Data que permite dar coherencia y orden al monitoreo de metadatos. Los divide bajo unos patrones asignados por el analista para ofrecer un panorama exacto de lo que se está investigando en un país concreto. Es decir, permite seleccionar un país concreto y visualizar el volumen de metadatos de comunicaciones que se han recogido y almacenado

Proporciona información muy detallada, por ejemplo, cuantificando de manera exacta todas las llamadas telefónicas y todos los e-mails recogidos y almacenados cada día en el mundo. Tal y como se puede observar en una de las diapositivas filtradas[79] sobre el programa, básicamente, proporciona respuestas a las siguientes preguntas:

- ¿Cuántos registros ha recolectado una unidad o un país en un período de tiempo determinado?
- ¿Existe alguna tendencia visible?
- ¿Qué activos se recogen de un país concreto y de qué tipo?
- ¿Cuál es el campo de visión de un sitio en concreto y de qué tipo?

[79] Snowden Doc Search. *Boundless Informant – Describing Mission Capabilities from Metadata Records* [en línea]. Journalistic Source Protection Defence Fund, 2012 [Consulta: 2 agosto 2016]. Disponible en: <https://search.edwardsnowden.com/docs/BoundlessInformant%E2%80%93DescribingMissionCapabilitiesfromMetadataRecords2013-06-08nsadocs>.

En la siguiente diapositiva podemos hacernos una idea de cómo se ven los resultados del programa.

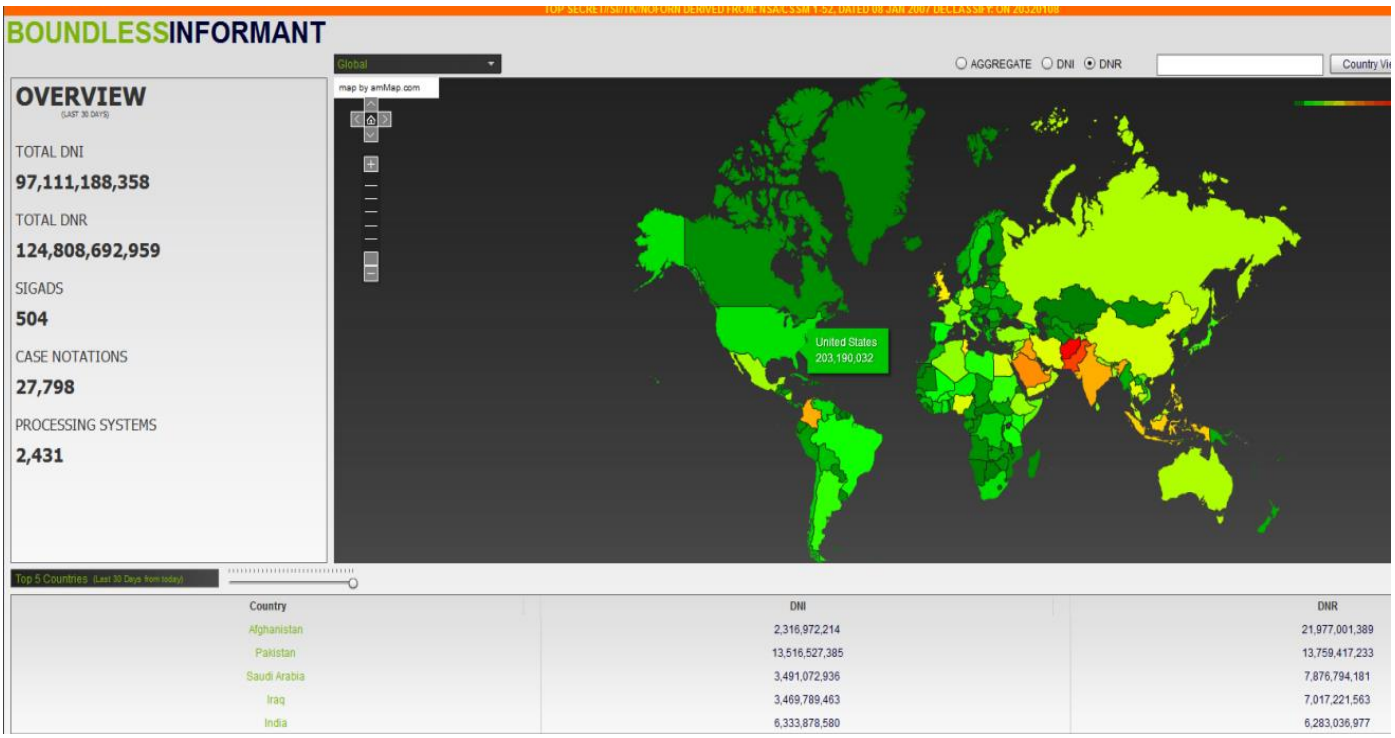

Tabla 46. Interfaz de Boundless Informant. (EdwardSnowden.com, 2015)

La imagen muestra la interfaz que vería cualquier persona que tuviera acceso al programa. Se despliega un mapa plano del mundo con los países pintados en varios colores, dependiendo de la cantidad de información captada, y detalla la cantidad de metadatos DNI y DNR recogidos en cada país.

Unified Targeting Tool (UTT)

Este software analítico se encarga de seleccionar objetivos concretos para su vigilancia, mayormente personas físicas concretas. Forma parte del sistema Turbulence, una potente arma cibernética de la NSA basada en la inyección de malwares. Para ejercer su tarea, contrasta los datos recopilados que entran al sistema a través de los malwares de Turbulence con una base de datos denominada Nymrod, que almacena nombres de personas y empresas.

Según las diapositivas filtradas sobre UTT, el programa puede filtrar la información por nacionalidad, localización y “extensión” (por ejemplo, diplomático). Dispone de una categoría denominada “inteligencia del propósito de la información”, donde el analista puede especificar un área geopolítica, un tema y un subtema.

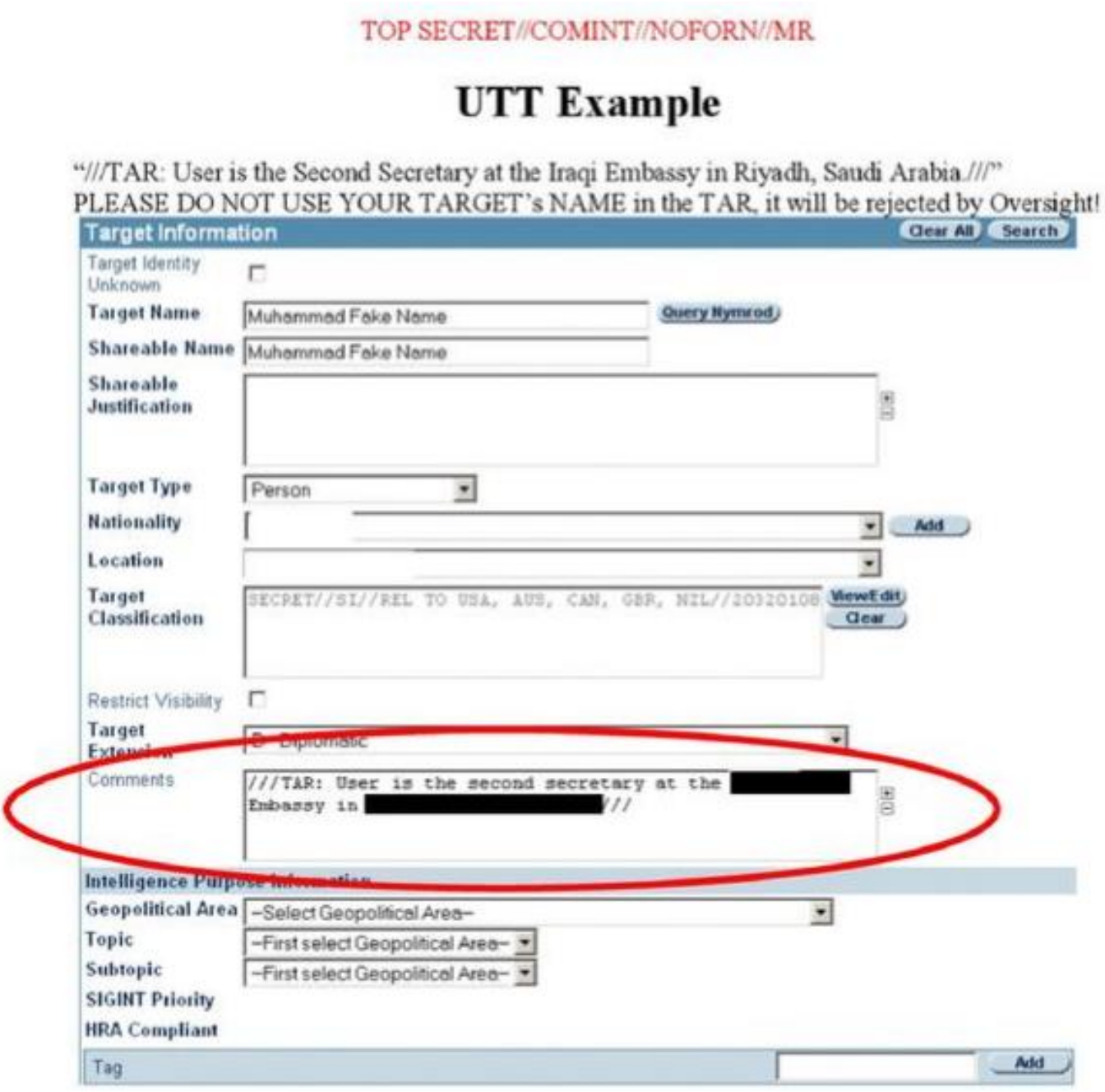

Tabla 47. Interfaz del buscador de UTT. (EdwardSnowden.com, 2015)

UTT da la opción de establecer la frecuencia con la que se transmite nueva información sobre la persona vigilada. Se puede observar la existencia de un campo denominado "Special Authorization", que seguramente sirva para vigilar a aquellos objetivos que requieran una orden judicial, como ciudadanos estadounidenses o extranjeros residentes en Estados Unidos.

Además, por lo que parece, el software puede atender consultas sin la necesidad de que el nombre del objetivo sea conocido. Para ello, podemos ver que el analista debe seleccionar un propósito, una fuente y un "factor extranjero", como los que aparecen desplegados en la captura de imagen.

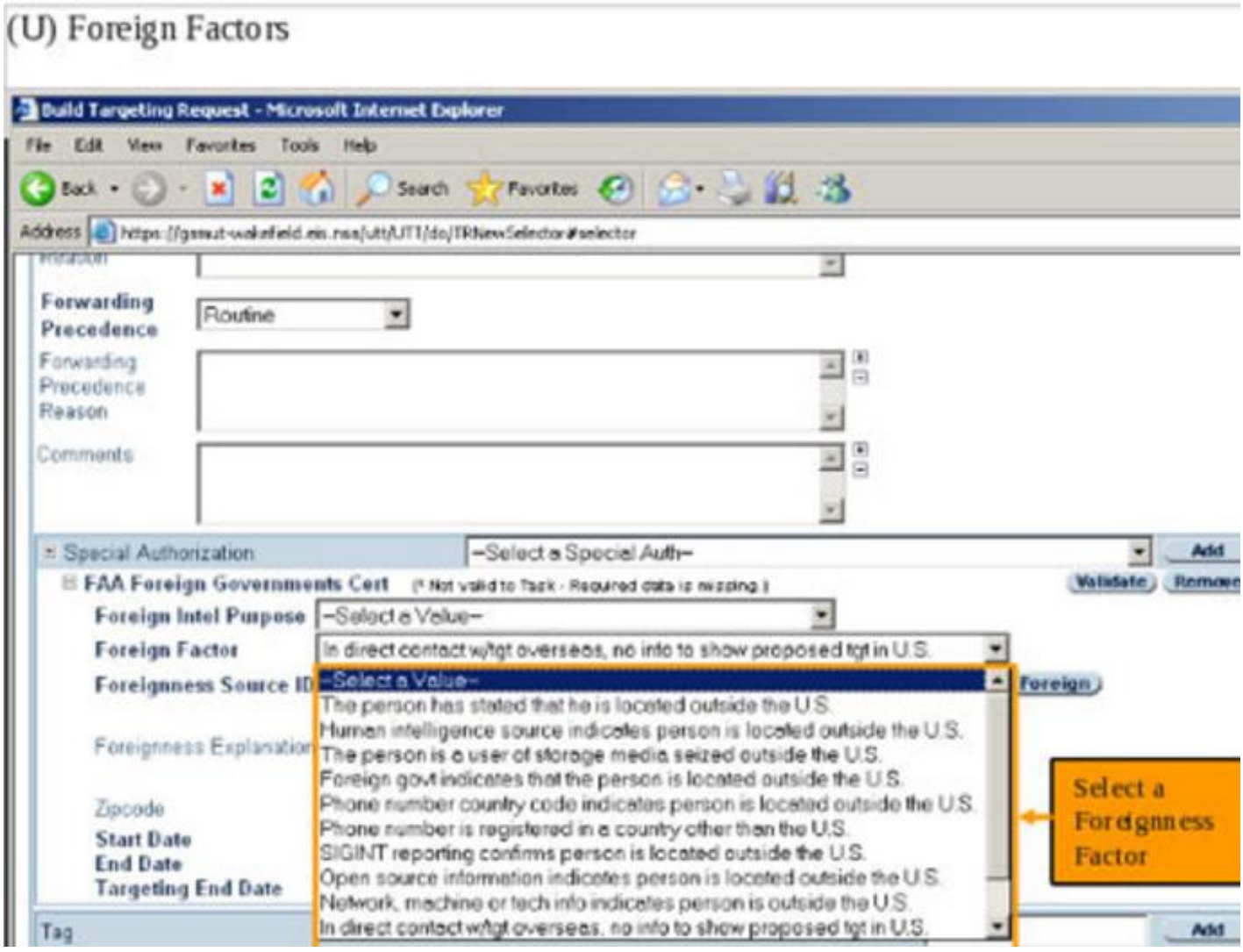

Tabla 48. Interfaz del buscador de UTT. (EdwardSnowden.com, 2015)

Accumulo

Accumulo es una base de datos NoSQL del tipo key/value y orientada en columnas. Es parecida a BigTable, la enorme base de datos de Google, pero con un nivel de seguridad mucho mayor y con más prestaciones. Es capaz de categorizar todos y cada uno de los datos que recibe, además de encontrar conexiones entre toda clase de datos aparentemente no relacionados. Agrupa conjuntos de datos relacionados entre sí con gran eficiencia, permitiendo descubrir información útil que de otra manera habría sido prácticamente imposible.

El sistema aprende automáticamente (machine learning), cuenta con un potente software de procesamiento de lenguaje natural y se aloja en una nube de escalabilidad horizontal. ¿Qué quiere decir todo esto? Quiere decir que Accumulo puede identificar patrones complejos entre los datos y predecir comportamientos futuros, realizar análisis de sentimientos para determinar reacciones y que está alojado en un sistema que cuantos más datos aloja más aumenta su rendimiento.

Como ya he dicho, es muy parecida en su estructura a la BigTable de Google, que está orientada a columnas, pero mientras que ésta solamente provee almacenamiento, la NSA necesita además procesar, analizar y establecer correlaciones de todos los datos que recopila, por lo que Accumulo cuenta con una serie de prestaciones extra.

Para ello, la NSA contó con la colaboración de Apache y su tecnología de código abierto Hadoop, que permite que la base de datos disponga de aún de más escalabilidad y de mucha más seguridad, hasta el punto de controlar accesos a nivel de celda.

Para hacernos una idea de su grado de escalabilidad, según Ely Kahn[80], el cofundador de Sqrrl Enterprise, la empresa que comercializa Accumulo, en la conocida base de datos orientada a columnas HBase, de la que ya hemos hablado en este trabajo, la escalabilidad disminuía al superar unos cuantos cientos de nodos. En Accumulo, sin embargo, la escalabilidad se mantiene con miles de nodos.

[80] Marko, Kurt. *The NSA and big data: what it can learn* [en línea]. Information Week, 2013 [Consulta: 10 agosto 2016]. Disponible en: <http://www.informationweek.com/big-data/big-data-analytics/the-nsa-and-big-data-what-it-can-learn/d/d-id/1110818?>.

Una de las piezas más importantes de esta base de datos es su aprendizaje automático y su capacidad de procesar lenguaje natural. Accumulo es capaz de re-adaptarse constantemente, re-ajustarse automáticamente para mejorar su rendimiento, actualizar consultas de búsqueda, interpretar datos o frases ambiguas e identificar objetos en imágenes digitales. Por ejemplo, Accumulo puede aplicar un contexto en el análisis textual de una frase ambigua como “esto lo va a destruir”, y determinar si se refiere a un plan de asesinato o a la opinión de un crítico sobre una novela.

Otras características para el análisis consisten en ofrecer a los analistas una interfaz amigable propia de una base de datos SQL, interfaces especiales para analizar estadísticas y múltiples sistemas de búsqueda, entre los que se incluye la búsqueda y recuperación de grafos[81], una herramienta de investigación muy valiosa para la NSA, tal y como mostró la propia agencia en una conferencia[82] en la Universidad de Carnegie Mellon en octubre de 2013. Para hacernos una idea de su capacidad en este sentido, Accumulo recoge más de 70 trillones de bordes (relaciones entre nodos) entre más de 4 trillones de nodos.

Es importante destacar que Accumulo se puede adquirir a través de Sqrrl Enterprise, por lo que es muy probable que otras agencias de inteligencia utilicen también este sistema adaptado a sus necesidades, como la NSA.

[81] Consultar la página 46 para ver más información sobre las bases de datos de grafos.
[82] Burkhardt, Paul; Waring, Chris. *An NSA Big Graph experiment* [en línea]. U.S. National Security Agency (NSA), 2013 [Consulta: 12 agosto 2016]. Disponible en: <http://www.pdl.cmu.edu/SDI/2013/slides/big_graph_nsa_rd_2013_56002v1.pdf>.

4.5 Explotación de los datos

El principal argumento de las agencias de inteligencia a la hora de justificar su acceso a la privacidad de los datos de la ciudadanía global es la seguridad nacional, es decir, garantizar la paz y la estabilidad de un país combatiendo amenazas tales como el terrorismo o el narcotráfico. El debate precisamente reside en este punto, ¿es necesario sacrificar la intimidad y la privacidad de las personas para proteger la seguridad nacional de un país?

Tanto la opinión pública como varias autoridades gubernamentales europeas consideran que la justificación de las agencias basadas en la seguridad nacional es una excusa. Sin ir más lejos, el Parlamento Europeo, a través del Informe Moraes, establece que este tipo de vigilancia masiva es inaceptable al ser totalmente indiscriminada y no basada en sospechas. El informe incluso apunta más lejos al sugerir "la posible existencia de otros fines ajenos al lucha antiterrorista" como "el espionaje político y económico".

De hecho, a la NSA cada vez le cuesta más sostener el argumento de la defensa nacional, ya que no se han presentado pruebas sobre alguno de los 54 atentados terroristas que dice haber prevenido gracias a los programas de vigilancia masiva. El medio norteamericano sin ánimo de lucro ProPublica[83] demostró que no había ninguna evidencia para sostener tal cosa.

En marzo de 2014, la Comisión LIBE estudió un documento[84] remitido por Edward Snowden a modo de respuesta ante la petición de la Comisión de dar testimonio en el caso de la vigilancia masiva electrónica. En él se explica que la vigilancia masiva es prácticamente ineficaz a la hora de combatir el terrorismo – de hecho, ha sido incapaz de prever los últimos ataques terroristas en Europa y Estados Unidos estos últimos años – y que sus principales objetivos son el espionaje económico, la vigilancia diplomática y el control social.

[83] Elliott, Justin; Meyer, Theodoric. *Claim on "Attacks Thwarted" by NSA spreads despite lack of evidence* [en línea]. ProPublica, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.propublica.org/article/claim-on-attacks-thwarted-by-nsa-spreads-despite-lack-of-evidence>.

[84] Snowden, Edward. *Introductory Statement* [en línea]. Comisión de Libertades Civiles, Justicia y Asuntos de Interior (LIBE), Parlamento Europeo, 2014 [Consulta: 13 agosto 2016]. Disponible en: <http://www.europarl.europa.eu/document/activities/cont/201403/20140307ATT80674/20140307ATT80674EN.pdf>.

Así pues, podríamos resumir que, en términos generales, la explotación de los datos procedentes de las actividades de vigilancia masiva de las agencias de inteligencia responde a los siguientes fines:

- Espionaje económico e industrial
- Espionaje político y diplomático
- Control social

He aquí algunos ejemplos que ilustran tales fines:

Ministerio de Minas y Energía de Brasil: El canal de televisión Globo reveló en octubre de 2013[85] una presentación de la agencia de inteligencia canadiense CSE que muestra un esquema detallado de las comunicaciones del Ministerio de Minas y Energía de Brasil, incluidas llamadas telefónicas, correos electrónicos y navegación en internet. Según Globo, el documento fue filtrado por Edward Snowden, quien lo obtuvo de una reunión en junio de 2012 de analistas de las agencias de inteligencia pertenecientes a los Cinco Ojos. Dilma Rousseff, la presidenta de Brasil, afirma que este hecho responde sin lugar a dudas a un acto de espionaje industrial.

Huawei: Según el New York Times[86], la NSA penetró en los servidores de la compañía de telefonía china Huawei, obteniendo información sobre las operaciones de la empresa y controlando las comunicaciones de sus directivos. Uno de sus objetivos era tratar de localizar vínculos entre la empresa y el Ejército Popular de Liberación Chino, pero el interés real era conocer cómo explotar la tecnología de Huawei para poder controlar comunicaciones de sus aparatos exportados a diferentes países. "Muchos de nuestros objetivos se comunican con productos producidos por Huawei", señala el documento de la NSA citado por el diario, "queremos asegurarnos de que sabemos cómo explotar estos productos para lograr acceso a redes de interés"

[85] The Guardian. *Brazil accuses Canada of spying after NSA leaks* [en línea]. The Guardian, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.theguardian.com/world/2013/oct/08/brazil-accuses-canada-spying-nsa-leaks>.

[86] Sanger, David; Perlroth, Nicole. *NSA breached chinese servers seen as security threat* [en línea]. The New York Times, 2014 [Consulta: 13 agosto 2016]. Disponible en: <http://www.nytimes.com/2014/03/23/world/asia/nsa-breached-chinese-servers-seen-as-spy-peril.html?partner=rss&emc=rss&smid=tw-nytimes&_r=0>.

Gobierno y economía de Japón: En julio de 2015 Wikileaks[87] reveló que la NSA vigila de cerca desde el año 2006 al gobierno japonés y a varias empresas niponas. Los informes filtrados muestran que la agencia se había asegurado un acceso profundo al funcionamiento interno del gobierno japonés para obtener de forma rutinaria información altamente sensible, como las relaciones del gobierno con Estados Unidos, cuestiones comerciales o posturas y políticas acerca del cambio climático. Según Wikileaks, la NSA espía a conglomerados japoneses, funcionarios del gobierno, ministros y asesores de alto rango, y tiene especial interés en los funcionarios del Banco Central de Japón, en el ministro de Economía, Comercio e Industria, en la división de gas natural de Mitsubishi y en la división de petróleo de Mitsui. La filtración demuestra que Estados Unidos también tiene acceso a cuestiones como las importaciones agrícolas y las disputas comerciales, planes de desarrollo de técnicas japonesas o políticas nucleares.

Líderes políticos: los medios de comunicación calculan que han sido más de 35 los dirigentes internacionales espiados, aunque destacan especialmente los casos de la canciller alemana Angela Merkel, del ex presidente venezolano Hugo Chávez y de la presidenta brasileña Dilma Rousseff. Al igual que en el caso japonés, la NSA puede acceder a información altamente sensible y confidencial, tanto política como económica.

Red TOR: Según el The Guardian[88], la NSA y la agencia inglesa GCHQ han utilizado diversos programas para atacar a la red anónima TOR con la finalidad de conocer la identidad de sus usuarios. La red TOR es sumamente compleja y muy segura, por lo que las agencias son incapaces de atacar a usuarios específicos. Según parece, la estrategia consiste en poner "trampas" para ver quién cae. Si bien es cierto que en TOR hay potenciales objetivos sospechosos de terrorismo, narcotráfico o hacking, gran parte de los usuarios son activistas, periodistas que utilizan la red para comunicarse con sus fuentes confidenciales de forma segura, abogados y ciudadanos que simplemente desean mantener su privacidad intacta.

---

[87] WikiLeaks. *Target Tokyo* [en línea]. WikiLeaks, 2015 [Consulta: 13 agosto 2016]. Disponible en: <https://wikileaks.org/nsa-japan/>.

[88] Schneier, Bruce. *Attacking TOR: how the NSA targets users' online anonymity* [en línea]. The Guardian, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.theguardian.com/world/2013/oct/04/tor-attacks-nsa-users-online-anonymity>.

ONU: Desde 2010 la ONU ha sido objeto repetitivo de la vigilancia ejercida por la NSA. Ese mismo año Wikileaks filtró una serie de documentación[89] de la agencia donde quedaba patente el control sobre varios líderes de la organización con la finalidad de obtener información sobre misiones y planes estratégicos. En el verano de 2012 la NSA se introdujo en el sistema de conferencias de la ONU e intervino unas 450 comunicaciones durante tres semanas. En febrero de 2016, Wikileaks publicó más documentación secreta[90] de la agencia, demostrando que el secretario general de la ONU, Ban Ki-Moon, había sido espiado durante sus reuniones privadas sobre el cambio climático.

Petrobras: Según el medio brasileño Globo[91], la red interna de la mayor compañía petrolera de Sudamérica, Petrobras, estuvo bajo la atenta vigilancia de la NSA durante el 2012. Según Globo, son varios los motivos para querer espiar al gigante petrolífero: información sobre nuevas reservas petrolíferas en el mundo, análisis sísmicos, técnicas y tecnologías de extracción, exploraciones de nuevas zonas, localización de depósitos altamente ricos en petróleo…

Desacreditación y manipulación de información on-line: en febrero de 2014, el periodista Glenn Greenwald[92] revelaba cómo a partir de su acceso a toda la documentación filtrada por Snowden, las agencias de inteligencia occidentales, especialmente las norteamericanas y las inglesas, se dedican a inyectar toda clase de información falsa en Internet para desacreditar y destruir la reputación de ciertas personas y manipular discursos activistas y políticos considerados poco deseables para los intereses de sus gobiernos. Entre las técnicas más destacadas para conseguirlo destacan: postear material falso y atribuirlo falsamente a una persona o un colectivo, hacerse pasar por víctimas de algún acto del objetivo cuya reputación quieren socavar, postear información negativa en foros generales y especializados y colgar material controvertido o de carácter sexual sobre alguien o sobre algún colectivo u organización.

---

[89] WikiLeaks. *Public Library of U.S. Diplomacy* [en línea]. WikiLeaks, 2010 [Consulta: 14 agosto 2016]. Disponible en: <https://wikileaks.org/plusd/cables/09STATE80163_a.html#efmJZLJeM>.

[90] WikiLeaks. *NSA targets world leaders for U.S. geopolitical interests* [en línea]. WikiLeaks, 2016 [Consulta: 14 agosto 2016]. Disponible en: <https://wikileaks.org/nsa-201602/>.

[91] Globo. *NSA documents show United States spied brazilian oil giant* [en línea]. Grupo Globo, 2013 [Consulta: 14 agosto 2016]. Disponible en: <http://g1.globo.com/fantastico/noticia/2013/09/nsa-documents-show-united-states-spied-brazilian-oil-giant.html>.

[92] Greenwald, Glenn. *How covert agents infiltrate the Internet to manipulate, deceive and destroy reputations* [en línea]. The Intercept, 2014 [Consulta: 14 agosto 2016]. Disponible en: <https://theintercept.com/2014/02/24/jtrig-manipulation/>.

Pero ya que este trabajo parte de la premisa de que las agencias de inteligencia son capaces de controlar y manipular – además de vigilar – a grandes masas de población, enfoquémonos en la cuestión del control social.

Para empezar, partimos de la base de que para justificar la vigilancia masiva, las agencias y los gobiernos han recurrido multitud de veces al lema de “si no tienes nada que ocultar no tienes nada que temer”. Ese enfoque viene a decir que cualquier persona que defienda y exija respeto por su privacidad está cometiendo actos reprochables, y si no, véase la obsesión de las agencias con desenmascarar a los usuarios anónimos de TOR.

Tras las filtraciones de documentos secretos por parte de Edward Snowden entorno a la vigilancia masiva – las más mediáticas de la historia sin lugar a dudas –, la población mundial fue realmente consciente de que la privacidad es hoy en día una ilusión. Desde entonces, en lugar de endurecer las medidas de protección de datos y de limitar las capacidades de las agencias de inteligencia, lo que han hecho la mayoría de gobiernos ha sido promulgar leyes que les dan aún más poder, tal y como veremos en el próximo apartado sobre la legislación relacionada.

Como hemos visto durante los últimos años, la seguridad nacional es una excusa para enmascarar al verdadero objetivo de gobiernos y agencias: conseguir más poder a través del control total de los ciudadanos. Las acciones de vigilancia a escala mundial sin ningún tipo de discriminación es algo propio de un Estado policial electrónico. La mayoría de la ciudadanía es consciente de ello, es consciente de que las actividades que están llevando a cabo a través de Internet, lo que están comprando o lo que están hablando a través de chats y llamadas puede estar siendo monitoreado y vigilado en ese mismo instante, a pesar de no estar relacionado con atentados, actos terroristas o narcotráfico.

El hecho de que un ciudadano sea consciente de que puede estar siendo vigilado constantemente recibe el nombre de panoptismo. El **panóptico** es un sistema carcelario donde cada prisionero es constantemente visible a los ojos de quien observa, mientras que desde su posición, el reo no puede saber quién lo observa, si es que lo observa alguien. La eficacia del panoptismo reside en ver sin ser visto, algo fundamental en todo sistema de vigilancia.

En la sociedad actual, la multiplicación y la complejidad de las relaciones entre individuos puede atentar contra el normal desarrollo y convivencia de las personas ubicadas en el sistema, por lo que el objetivo final es garantizar el orden social.

Es en este punto donde el esquema de poder disciplinario propuesto por el panóptico cobra importancia, debido a que sus mecanismos de observación son capaces de modificar comportamientos y conductas, asegurando con ello el orden y la adhesión social. El avance tecnológico perpetúa este sistema, ya que otorga herramientas de vigilancia cada vez más sofisticadas.

La vigilancia masiva genera miedo y autocensura en la sociedad, tal y como confirma un reciente estudio de la Universidad de Oxford[93]. Su sola existencia genera miedo y asfixia a la libertad de expresión. Por ejemplo, tras las revelaciones de Snowden hubo un declive del 20% en las visitas a artículos de la Wikipedia relacionados con el terrorismo. De acuerdo con estos resultados, las revelaciones habrían afectado la manera en la que las personas interactúan con eventos significativos, como la guerra, y por ende, el estado de la libertad de expresión en general y el diálogo democrático. Al igual que en el panóptico, las personas vigiladas alteran su comportamiento ante la expectativa de estar siendo observadas en cualquier momento, sin tener ninguna manera de saber a ciencia cierta cuándo están siendo vigiladas de verdad.

Otro dato revelador del estudio es que los musulmanes que han participado en el proyecto no creen haber alterado su comportamiento, a pesar de que el 71,7% creen que el Gobierno estadounidense está vigilando sus actividades. La evidencia empírica muestra que, aunque pensemos que nuestro comportamiento no se ha visto alterado por el miedo y la autocensura, la realidad es otra.

La vigilancia masiva empuja a la sociedad a convertirse en sus propios censores, ya que uno no puede tener la certeza de cuáles de sus acciones podrían ser consideradas sospechosas. Si a esto le sumamos la sensación de estar siendo constantemente vigilados, la autocensura se hará presente también en situaciones íntimas, cotidianas y domésticas, ya que una simple búsqueda en Google podría poner a un individuo en la lista de personas problemáticas o sospechosas. Nada de esto tiene sentido en países supuestamente democráticos que se jactan de ofrecer a sus ciudadanos el derecho a la libre expresión, al debate, a la participación política o al acceso a la información.

[93] Penney, Jon. *Chilling effects: online surveillance and Wikipedia use* [en línea]. Oxford Internet Institute, University of Oxford, 2016 [Consulta: 15 agosto 2016]. Disponible en: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645>.

5 Legislación

El apartado sobre legislación es fundamental para comprender cómo han reaccionado ante el escándalo de la vigilancia masiva los órganos de poder público de los Estados Unidos y de la Unión Europea y hacia dónde van encaminados los estados democráticos en esta cuestión.

Desde la publicación en la prensa mundial de los documentos filtrados por el ex analista de la CIA y de la NSA, Edward Snowden, en junio de 2013, hasta la actualidad, el contexto legal en el que se enmarca el trabajo ha ido variando a lo largo de estos tres años.

En primer lugar destaca la ley norteamericana conocida como **Patriot Act**, una ley aprobada tras los atentados del 11-S que amplía la capacidad de control del Estado y la de sus agencias de inteligencia. La Patriot Act va en contra de los derechos constitucionales norteamericanos al atentar contra los derechos humanos y las libertades civiles, y es la encargada de dar cobertura legal a las actividades de vigilancia masiva de la NSA tanto dentro de los Estados Unidos como en suelos extranjeros.

Bajo la estadounidense Ley de Vigilancia de la Inteligencia Extranjera **(FISA)**, el Tribunal de Vigilancia de Inteligencia Extranjera de los Estados Unidos de América **(FISC)** es el encargado de supervisar las solicitudes de vigilancia hechas por las agencias de seguridad federales contra sospechosos extranjeros que se encuentren dentro de los Estados Unidos. Todas las solicitudes de las órdenes de vigilancia (conocidas como órdenes FISA) son presentadas ante un juez individual del tribunal. Debido a la naturaleza sensible de su actividad, el tribunal es un "tribunal secreto", de manera que sus audiencias están cerradas al público.

En junio de 2013 se filtró una orden emitida por el FISC del 25 de abril de 2013, donde se instaba a la compañía de telecomunicaciones Verizon a entregar a la NSA todos los registros, estadísticas globales y datos de geolocalización de todas las llamadas registradas en su sistema, incluyendo las llamadas telefónicas locales de ciudadanos estadounidenses. Esta filtración fue la primera de muchas publicada en los medios de comunicación internacionales que destaparían la trama del espionaje masivo a lo largo del año 2013.

En Europa, la trama del espionaje masivo supone una grave vulneración del Convenio Europeo de Derechos Humanos **(CEDH)**, un convenio basado en la Declaración Universal de los Derechos Humanos y adoptado por el Consejo de Europa en 1950. En virtud del CEDH, todos los ciudadanos europeos tienen derecho a ser protegidos de injerencias arbitrarias o ilegales en su vida privada, su familia, su domicilio o sus comunicaciones, así como de ataques ilegales a su honra y reputación. Este derecho debe estar garantizado sin importar de dónde provengan tales injerencias (autoridades estatales o personas físicas o jurídicas).

El 4 de julio de 2013 el Parlamento Europeo (PE) aprueba una resolución encargando a su Comisión de Libertades Civiles, Justicia y Asuntos de Interior (LIBE) una investigación exhaustiva de los programas de vigilancia masiva.

El 18 de diciembre de ese mismo año, la Asamblea General de la ONU aprueba la **Resolución 68/167**[94] sobre el derecho a la privacidad en la era digital, copatrocinada por 57 Estados Miembros, donde se expone que los derechos de los ciudadanos deben estar protegidos en Internet, y que para ello hace falta que los Estados respeten y protejan el derecho a la privacidad en las comunicaciones digitales. La resolución afirma que los Estados deben examinar sus respectivas legislaciones sobre vigilancia, interceptación y recopilación de datos personales para cambiar sus prácticas y procedimientos.

El 21 de febrero de 2014 el LIBE presenta al Parlamento Europeo los resultados de la investigación sobre los programas de vigilancia masiva bajo el denominado **Informe Moraes**. En las recomendaciones del informe, el LIBE pide lo mismo que la ONU: que los Estados Miembros de la UE prohíban las actividades de vigilancia masiva generalizada y que se aseguren de que todos sus marcos legislativos y mecanismos de control actuales y futuros por los que se rigen las actividades de los servicios de inteligencia se atengan a los estándares del CEDH y a la legislación en materia de protección de datos de la Unión Europea.

[94] Asamblea General de las Naciones Unidas. *Resolución aprobada por la Asamblea General el 18 de diciembre de 2013. 68/167. El derecho a la privacidad en la era digital* [en línea]. Organización de las Naciones Unidas, 2013 [Consulta: 3 agosto 2016]. Disponible en: <http://www.un.org/es/comun/docs/?symbol=A/RES/68/167>.

El 8 de abril de 2014, El Tribunal de Justicia de la Unión Europea dicta una sentencia donde declara inválida la **Directiva 2006/24/CE** sobre la conservación de datos, que establecía la obligación de conservar durante un tiempo determinado un número considerable de datos generados o tratados en el marco de las comunicaciones electrónicas efectuadas por los ciudadanos en todo el territorio de la Unión Europea.

El Tribunal de Justicia de la Unión Europea, en su sentencia del 6 de octubre de 2015, declara no válido el tratado de **Puerto Seguro** de los Estados Unidos por no ofrecer una adecuada protección de los datos personales a los ciudadanos de la UE.

El tratado de Puerto Seguro es un mecanismo que establece un nivel de protección que la Unión Europea considera adecuado para las transferencias internacionales de datos a Estados Unidos. La sentencia proclama que la decisión de Puerto Seguro es inválida por dos motivos:

- Porque entiende que prevalece incondicionalmente y sin ninguna limitación la seguridad nacional, el interés público y el cumplimiento de la ley sobre los derechos fundamentales a la intimidad y la protección de datos, sin otorgar a los ciudadanos europeos ningún medio para obtener la tutela efectiva de esos derechos.
- Porque no otorga a los Estados Miembros un margen suficiente para suspender las transferencias en caso de que estos apreciaran una vulneración de los derechos de los ciudadanos europeos.

Curiosamente, el único país implicado que limita el poder de sus agencias de inteligencia es Estados Unidos. El 2 de junio de 2015 se aprueba la **USA Freedom Act**, derogando a la Patriot Act. Un elemento central de la ley es que retira a la NSA la capacidad de almacenar los datos sobre las llamadas telefónicas de millones de estadounidenses y coloca estos datos en manos de las compañías telefónicas.

Los espías podrán acceder a estos datos solamente caso a caso y con previa autorización judicial. Sin embargo, cabe destacar que las nuevas resoluciones se centran únicamente en los ciudadanos estadounidenses, por lo que la situación legislativa estadounidense sobre ciudadanos europeos continúa exactamente igual.

A pesar de las anteriores resoluciones de la ONU y del Parlamento Europeo, tal y como podemos observar en la **Resolución 2015/2635**[95] del Parlamento Europeo en octubre de 2015, los Estados Miembros no muestran ningún tipo de sentido de la urgencia ni de voluntad a la hora de abordar las cuestiones planteadas por la ONU y por el propio Parlamento Europeo, así como la inexistencia de haber aplicado alguna de las recomendaciones específicas de la anterior resolución europea.

En la misma resolución, el Parlamento Europeo expresa su preocupación por algunas de las leyes recientes de algunos Estados Miembros que amplían las capacidades de vigilancia de sus agencias de inteligencia, entre las que destacan:

- **Ley sobre los servicios de inteligencia franceses** adoptada y aprobada el 24 de junio de 2015 por la Asamblea Nacional francesa. La nueva ley contempla la vigilancia e interceptación de información y comunicaciones en la red y en teléfonos móviles, la instalación en vehículos de mecanismos de localización por GPS y la colocación de micrófonos y cámaras. No son necesarias órdenes judiciales previas, pero para solventar ese hecho la ley crea un nuevo organismo, la Comisión Nacional de Control de las Técnicas de Información (CNCTR), que será la encargada de verificar que la inteligencia francesa no atente contra las libertades públicas.

- En el Reino Unido, en vista de la anulación de la Directiva 2006/24/CE, se aprueba en el 2014 la **Ley de Retención de Datos y del Poder Investigador (DRIPA)**, por la cual los servicios de inteligencia ingleses tienen la capacidad de obligar a las empresas de Internet y de telecomunicaciones a retener y conservar sus datos durante un cierto periodo de tiempo.

[95] Parlamento Europeo. *Resolución del Parlamento Europeo, de 29 de octubre de 2015, sobre el seguimiento de la Resolución del Parlamento Europeo, de 12 de marzo de 2014, relativa a la vigilancia electrónica masiva de los ciudadanos de la UE (2015/2635 RSP)* [en línea]. Estrasburgo: Parlamento Europeo, 2015. [Consulta: 4 agosto 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2015-0388+0+DOC+XML+V0//ES>.

El 27 de abril de 2016 sale a la luz, tras cuatro años de preparación, la nueva ley de protección de datos europea: **Reglamento (UE) 2016/679**[96] del Parlamento Europeo y del Consejo relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

El objetivo del nuevo reglamento general es dar más control a los ciudadanos sobre su información privada en un mundo de teléfonos inteligentes, redes sociales, banca por internet y transferencias globales.

El nuevo paquete de protección de datos también incluye una directiva sobre transmisión de datos para cuestiones judiciales y policiales. Se aplica al intercambio de datos transfronterizos dentro de la Unión Europea y establece unos estándares mínimos para el tratamiento de datos en cada país.

Finalmente, en julio de 2016, tras meses de negociaciones, la Unión Europea y los Estados Unidos ratifican el acuerdo **Privacy Shield**, que llena el hueco dejado por el tratado de Puerto Seguro. El nuevo marco legal para proteger los datos personales de los ciudadanos europeos que sean transferidos a suelo estadounidense aspira a proteger los datos personales de los europeos y proporcionar seguridad jurídica para las empresas, así como recuperar la confianza de los consumidores en el contexto de las transferencias transatlánticas de datos.

[96] Parlamento Europeo; Consejo de la Unión Europea. *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)* [en línea]. Diario Oficial de la Unión Europea, 2016 [Consulta: 4 agosto 2016]. Disponible en: <https://www.boe.es/doue/2016/119/L00001-00088.pdf>.

6 Conclusiones

Una vez finalizado el trabajo he podido llegar a una serie de conclusiones que responden tanto a los objetivos como a las hipótesis formuladas al principio del estudio.

En primer lugar, tal y como hemos podido ver en el apartado de conceptos introductorios, un data center puede ser considerado como un archivo. Al igual que en un archivo, los datos de un data center son recibidos y producidos por personas físicas y jurídicas, públicas o privadas, y son fruto de sus actividades. Así mismo, un data center es a su vez una institución que gestiona todos estos datos y un espacio físico donde éstos se conservan.

Por otro lado, en el apartado sobre la explotación de los datos por parte de las agencias de inteligencia, hemos podido observar cómo uno de sus objetivos principales es ejercer el control de la sociedad, por encima del objetivo de proteger la seguridad nacional. Este hecho se confirma al no haber ningún tipo de discriminación a la hora de recopilar datos de la ciudadanía, tratando a todo individuo como sospechoso, y al influir en la red desacreditando a personas, entidades, discursos y manipulando información on-line.

En cuanto a la hipótesis sobre que las tecnologías Big Data vulneran la privacidad de los datos de las personas, creo que su veracidad ha quedado patente a lo largo de todo el trabajo. Por una parte, las empresas tecnológicas recopilan datos procedentes de toda clase de actividades e interacciones que realicemos en la red bajo unos términos y condiciones de uso altamente cuestionables y a través de herramientas que en la mayoría de los casos no cumple con la actual ley europea sobre protección de datos. Por la otra, hemos podido ver cómo las agencias de inteligencia vulneran reiteradamente todas las normativas nacionales e internacionales sobre retención de datos, privacidad e intimidad.

La última hipótesis parte de que los gobiernos democráticos mencionados en este trabajo son partícipes de las prácticas de vigilancia masiva indiscriminada que realizan sus agencias de inteligencia. Queda patente este hecho solamente con analizar los informes emitidos por la Unión Europea al respecto. El informe Moraes confirma la existencia de alianzas entre Estados Unidos y varios Países Miembros, donde las agencias colaboran entre ellas intercambiando datos de ciudadanos y tecnologías de vigilancia.

Además, tras el escándalo de la vigilancia masiva en junio de 2013, muchos Países Miembros no solamente no actuaron en contra de ella, si no que modificaron su legislación nacional para dar aún más poder a sus agencias de inteligencia.

En lo relativo a los procesos documentales que se siguen en los data centers aquí analizados, se llega a la conclusión de que son perfectamente extrapolables a lo que entendemos por archivo digital. Si bien el origen de los datos y las tecnologías empleadas son distintos, en ambos se produce una entrada categorizada de datos, una gestión de ellos a través de una serie de normativas internas y protocolos y un almacenamiento con determinadas políticas de acceso, seguridad y conservación que permiten la recuperación de los datos almacenados. A pesar de ello, no se ha podido encontrar documentación que defina cómo son estas políticas, más allá de características como las de la base de datos de la NSA Accumulo, que ofrece niveles de seguridad y protección a nivel de bit.

Por otro lado, podemos determinar que las tecnologías utilizadas en los distintos procesos de gestión documental de un data center son los máximos exponentes tecnológicos del Big Data. En este sentido, destacan especialmente los programas de extracción, interceptación y recolección de datos de la NSA y de la agencia inglesa GCHQ, que como hemos podido ver, son extremadamente complejos y poseen un alcance mundial. De la misma forma, destacan bases de datos tan impactantes como Accumulo, cuya capacidad de almacenamiento y análisis es abrumadora, así como el sistema de relaciones e interacciones existentes entre las decenas de bases de datos de la NSA que permiten relacionar y encadenar datos dispares alojados en distintos sistemas.

Cabe destacar también que los data centers de las grandes empresas tecnológicas estudiadas son una de las principales fuentes de entrada de datos de los data centers de las agencias de inteligencia. Algunas de ellas están asociadas con las agencias, como es el caso de Microsoft y de las grandes compañías de telecomunicaciones, como Vodafone o BT, que dan acceso a las agencias a sus cables de fibra óptica. Sin embargo, no se puede saber con certeza si el resto de las compañías aquí analizadas son también proveedores conscientes, aunque hemos podido ver que hay pruebas sobre la creación de puertas traseras en sus sistemas de información.

Finalmente, me gustaría concluir el trabajo afirmando que las tecnologías Big Data, sin lugar a dudas, aportan múltiples beneficios a las organizaciones que sepan cómo gestionar estas herramientas correctamente. Los datos hablan, y hoy en día, si una empresa privada no hace uso de estos sistemas no puede aspirar a competir con otras empresas del sector que sí lo hacen.

Sabemos que Big Data ha supuesto una revolución en la gestión de los datos y que ha comportado la creación de nuevos puestos de trabajo y nuevas oportunidades para todo tipo de sectores en el mercado. Pero tal y como hemos podido ver en este trabajo, también sabemos que Big Data puede llegar a ser la tecnología más intrusiva de la historia y el responsable de la desaparición de la privacidad de los datos de los ciudadanos de todo el mundo.

7 Bibliografía

Acens. *Bases de datos NoSQL. Qué son y tipos que nos podemos encontrar* [en línea]. Telefónica, 2014 [Consulta: 29 abril 2016]. Disponible en: <https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf>.

Ackerman, Spencer. *U.S. tech giants knew of NSA data collection, agency's top lawyer insists* [en línea]. Washington: The Guardian, 2014 [Consulta: 2 julio 2016]. Disponible en: <https://www.theguardian.com/world/2014/mar/19/us-tech-giants-knew-nsa-data-collection-rajesh-de>.

Apache Software Fundation. *Hadoop* [en línea]. 2014. Última actualización: 11 febrero 2016. [Consulta: 20 marzo 2016]. Disponible en: <http://hadoop.apache.org/>.

Asamblea General de las Naciones Unidas. *Resolución aprobada por la Asamblea General el 18 de diciembre de 2013. 68/167. El derecho a la privacidad en la era digital* [en línea]. Organización de las Naciones Unidas, 2013 [Consulta: 3 agosto 2016]. Disponible en: <http://www.un.org/es/comun/docs/?symbol=A/RES/68/167>.

Ball, James. *NSA collects milions of text messages daily in "untargeted" global sweep* [en línea]. The Guardian, 2014 [Consulta: 4 julio 2016]. Disponible en: <https://www.theguardian.com/world/2014/jan/16/nsa-collects-millions-text-messages-daily-untargeted-global-sweep>.

Ball, James. *NSA stores metadata of millions of web users for up to a year, secret files show* [en línea]. The Guardian, 2013 [Consulta: 23 enero 2016]. Disponible en: <https://www.theguardian.com/world/2013/sep/30/nsa-americans-metadata-year-documents>.

Ball, James; Harding, Luke; Garside, Juliette. *BT and Vodafone among telecoms companies passin details to GCHQ* [en línea]. The Guardian, 2013 [Consulta: 18 julio 2016]. Disponible en: <https://www.theguardian.com/business/2013/aug/02/telecoms-bt-vodafone-cables-gchq>.

Bandera, Magda. Lo que el sistema sabe sobre ti. *Revista Playboy*. 2003. Núm. 3, época 2.

BBVA Open4U. *Bigtable, el servicio de base de datos NoSQL con el que Google quiere dominar los Big Data* [en línea]. BBVA, 2015 [Consulta: 8 abril 2016]. Disponible en: <http://www.bbvaopen4u.com/es/actualidad/bigtable-el-servicio-de-base-de-datos-nosql-con-el-que-google-quiere-dominar-los-big-data>.

BBVA Open4U. Qué es y para qué sirve una base de datos orientada a grafos [en línea]. BBVA, 2015 [Consulta: 8 abril 2016]. Disponible en: <https://bbvaopen4u.com/es/actualidad/neo4j-que-es-y-para-que-sirve-una-base-de-datos-orientada-grafos>.

Big data [en línea]. New York: Mary Ann Liebert, Inc. publishers, 2013-2016 [Consulta: 21 enero 2016]. Disponible en: <http://online.liebertpub.com/loi/big>. ISSN 2167-647X.

Bundesministerium der Finanzen. *Einzelpläne* [en línea]. 2015 [Consulta: 13 mayo 2016]. Disponible en: <http://www.bundeshaushalt-info.de/#/2015/soll/ausgaben/einzelplan/0404.html>.

Burkhardt, Paul; Waring, Chris. *An NSA Big Graph experiment* [en línea]. U.S. National Security Agency (NSA), 2013 [Consulta: 12 agosto 2016]. Disponible en: <http://www.pdl.cmu.edu/SDI/2013/slides/big_graph_nsa_rd_2013_56002v1.pdf>.

Caño, Antonio. *La NSA afirma que el espionaje masivo fue realizado por Francia y España* [en línea]. El País, 2015 [Consulta: 4 julio 2016]. Disponible en: <http://internacional.elpais.com/internacional/2013/10/29/actualidad/1383074305_352044.html>.

Comisión de Libertades Civiles, Justicia y Asuntos de Interior. *Informe sobre el programa de vigilancia de la Agencia Nacional de Seguridad de los EEUU, los órganos de vigilancia en diversos Estados miembros y su impacto en los derechos fundamentales de los ciudadanos de la UE y en la cooperación transatlántica en materia de Justicia y Asuntos de Interior* [en línea]. Parlamento Europeo, 2013 [Consulta: 24 abril 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A7-2014-0139+0+DOC+XML+V0//ES>.

Corera, Gordon. *Escándalo de espionaje: qué es el "Club de los cinco ojos"* [en línea]. BBC, 2013 [Consulta: 5 mayo 2016]. Disponible en: <http://www.bbc.com/mundo/noticias/2013/10/131030_internacional_estados_unidos_espionaje_reino_unido_club_cinco_ojos_az>.

Danish Security and Intelligence Service. *PETs finances* [en línea]. 2015 [Consulta: 13 mayo 2016]. Disponible en: <https://www.pet.dk/English/About%20PET/PETs%20finances.aspx>.

EdwardSnowden.com. *Snowden doc search* [en línea]. Journalistic Source Protection Defence Fund, 2013. Disponible en: <https://search.edwardsnowden.com/>.

El País. *Microsoft compra Skype por 5.920 millones de euros* [en línea]. Barcelona: Ediciones El País S.L., 2011 [Consulta: 24 febrero 2016]. Disponible en: <http://tecnologia.elpais.com/tecnologia/2011/05/10/actualidad/1305018061_850215.html>.

Elastic. *Ingest attachment processor Plugin* [en línea]. Elasticsearch, 2016 [Consulta: 15 julio 2016]. Disponible en: <https://www.elastic.co/guide/en/elasticsearch/plugins/master/ingest-attachment.html>.

Electrospaces. "Section 215 bulk telephone records and the MAINWAY database". Blog Electrospaces. 15 de febrero de 2016. Blog. Acceso el 17 de julio de 2016. Disponible en: <http://electrospaces.blogspot.com.es/2016/01/section-215-bulk-telephone-records-and.html>.

Elliott, Justin; Meyer, Theodoric. *Claim on "Attacks Thwarted" by NSA spreads despite lack of evidence* [en línea]. ProPublica, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.propublica.org/article/claim-on-attacks-thwarted-by-nsa-spreads-despite-lack-of-evidence>.

Europa Press. *AI alerta de que los gobiernos mantienen la vigilancia masiva tras el caso Snowden* [en línea]. Madrid: Europa Press, 2015 [Consulta: 1 agosto 2016]. Disponible en: <http://www.europapress.es/internacional/noticia-ai-alerta-gobiernos-mantienen-vigilancia-masiva-caso-snowden-20150605190216.html>.

FBI. *Mission & Priorities* [en línea]. U.S. Department of Justice, 2015 [Consulta: 17 mayo 2016]. Disponible en: <https://www.fbi.gov/about/mission>.

Ferrer-Sapena, Antonia; Sánchez Pérez, Enrique A. Open data, big data: ¿hacia dónde nos dirigimos? [en línea]. *Anuario ThinkEPI*, febrero 2013, v.7 [Consulta: 3 febrero 2016]. Disponible en: <http://www.thinkepi.net/open-data-big-data-hacia-donde-nos-dirigimos>.

Fidelity Worldwide Investment. *Big data: una "revolución industrial" en la gestión de los datos digitales* [en línea]. Fidelity Worldwide Investment, 2012 [Consulta: 17 febrero 2016]. Disponible en: <https://www.fondosfidelity.es/static/pdfs/informes-fondos/Fidelity_ArgInvSXXI_BigData_Sept12_ES.pdf>.

Follorou, Jacques. *Surveillance: la DGSE a transmis des données à la NSA américaine* [en línea]. Le Monde, 2013 [Consulta: 9 julio 2016]. Disponible en: <http://www.lemonde.fr/international/article/2013/10/30/surveillance-la-dgse-a-transmis-des-donnees-a-la-nsa-americaine_3505266_3210.html>.

Full session – Big Data's "janitor" problem – Is it killing ROI? 14 de mayo de 2015. Acceso el 2 de abril de 2016. Vídeo de Youtube. Disponible en: <https://www.youtube.com/watch?v=YL5TK0vbuhc&feature=youtu.be>.

¿Qué es el club de los cinco ojos? 22 de noviembre de 2013. Acceso el 5 de mayo de 2016. Vídeo de Youtube. Disponible en: <https://www.youtube.com/watch?v=3HZLBr1uxf8>.

García Huerta, Ana. *Big Data y su impacto en el negocio: Una aproximación al valor que el análisis extremo de datos aporta a las organizaciones* [en línea]. Madrid: Oracle, 2012, p.6 [Consulta: 17 febrero 2016]. Disponible en: <https://emeapressoffice.oracle.com/imagelibrary/downloadMedia.ashx?MediaDetailsID=2197>.

Gellman, Barton; Poitras, Laura. *US intelligence mining data from nine US internet companyies in broad secret program* [en línea]. The Washington Post, junio 2013 [Consulta: 5 febrero 2016]. Disponible en: <https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html>.

Gellman, Barton; Soltani, Ashkan. *NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say* [en línea]. The Washington Post, 2013 [Consulta: 15 julio 2016]. Disponible en: <https://www.washingtonpost.com/world/national-security/nsa-infiltrates-links-to-yahoo-google-data-centers-worldwide-snowden-documents-say/2013/10/30/e51d661e-4166-11e3-8b74-d89d714ca4dd_story.html>.

Giones-Valls, Aina. Cuantificarse para vivir a través de los datos: los datos masivos (big data) aplicados al ámbito personal. *BiD, textos universitaris de Biblioteconomia i Documentació* [en línea]. Junio de 2015, núm. 34 [Consulta: 17 junio 2016]. Disponible en: <http://bid.ub.edu/es/34/giones.htm>.

Globo. *NSA documents show United States spied brazilian oil giant* [en línea]. Grupo Globo, 2013 [Consulta: 14 agosto 2016]. Disponible en: <http://g1.globo.com/fantastico/noticia/2013/09/nsa-documents-show-united-states-spied-brazilian-oil-giant.html>.

Google. *Ubicaciones de los centros de datos* [en línea]. 2012 [Consulta: 8 agosto 2016]. Disponible en: <https://www.google.com/about/datacenters/inside/locations/index.html>.

Greenwald, Glenn. *How covert agents infiltrate the Internet to manipulate, deceive and destroy reputations* [en línea]. The Intercept, 2014 [Consulta: 14 agosto 2016]. Disponible en: <https://theintercept.com/2014/02/24/jtrig-manipulation/>.

Greenwald, Glenn. *Sin un lugar donde esconderse. Edward Snowden, la NSA y el Estado de vigilancia de EE.UU*. Nueva York: Metropolitan Books, 2014. ISBN 978-84-666-5459-3.

Greenwald, Glenn [et.al]. *Microsoft handed the NSA access to encrypted messages* [en línea]. The Guardian, 2013 [Consulta: 28 junio 2016]. Disponible en: <https://www.theguardian.com/world/2013/jul/11/microsoft-nsa-collaboration-user-data>.

Greenwald, Glenn; Aranda, Germán. *El CNI facilitó el espionaje masivo de EEUU a España* [en línea]. El Mundo, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://www.elmundo.es/espana/2013/10/30/5270985d63fd3d7d778b4576.html>.

Greenwald, Glenn; Ball, James; Borger, Julian. *Revealed: how US and UK spy agencies defeat Internet privacy and security* [en línea]. The Guardian, 2013 [Consulta: 21 junio 2016]. Disponible en: <https://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security>.

Greenwald, Glenn; MacAskill, Ewen. *NSA Prism program taps in to user data of Appe, Google and others* [en línea]. The Guardian, junio 2013 [Consulta: 5 febrero 2016]. Disponible en: <http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data>.

Hadoop Wiki. *Powered by Apache Hadoop* [en línea]. Última actualización 8 diciembre 2015 [Consulta: 23 febrero 2016]. Disponible en: <http://wiki.apache.org/hadoop/PoweredBy>.

Henschen, Doug. *Defending NSA Prism's Big Data tools* [en línea]. Information Week, 2013 [consulta: 8 agosto 2016]. Disponible en: <http://www.informationweek.com/big-data/big-data-analytics/defending-nsa-prisms-big-data-tools/d/d-id/1110318?>.

IBM Institute for Business Value. *Analytics: el uso de big data en el mundo real. Cómo las empresas más innovadoras extraen valor de datos inciertos* [en línea]. IBM Global Business Services, 2012, p. 6-7 [Consulta: 17 febrero 2016]. Disponible en: <http://www-05.ibm.com/services/es/bcs/pdf/Big_Data_ES.PDF>.

Joa, Carlos. *Consideraciones para el diseño y la construcción de un data center* [en línea]. Slideshare, 2015 [Consulta: 22 febrero 2016]. Disponible en: <http://es.slideshare.net/kacjoa/diseo-y-normas-para-data-centers>.

Jeffries, Adrianne. *Escape from Prism: how Twitter defies government data-sharing* [en línea]. The Verge, 2013 [Consulta: 2 marzo 2016]. Disponible en: <http://www.theverge.com/2013/6/13/4426420/twitter-prism-alex-macgillivray-NSA-government>.

Lakshman, Avinash. *Cassandra: a structured storage system on a P2P Network* [en línea]. Facebook Engineering, 2008 [consulta: 23 febrero 2016]. Disponible en: <https://www.facebook.com/note.php?note_id=24413138919&id=9445547199&index=9>.

Lobosco, Katie. *Google, Facebook… Paltalk?!* [en línea]. Nueva York: CNN, 2013 [Consulta: 24 febrero 2016]. Disponible en: <http://money.cnn.com/2013/06/07/technology/security/paltalk-nsa-surveillance/>.

MacAskill, Ewen [èt.al]. *GCHQ taps fibre-optic cables for secret access to world's communications* [en línea]. The Guardian, 2013 [Consulta: 6 junio 2016]. Disponible en:<https://www.theguardian.com/uk/2013/jun/21/gchq-cables-secret-world-communications-nsa>.

MacAskill, Ewen [èt.al]. *Mastering the Internet: how GCHQ set out to spy on the world wide web* [en línea]. The Guardian, 2013 [Consulta: 6 junio 2016]. Disponible en: <https://www.theguardian.com/uk/2013/jun/21/gchq-mastering-the-internet>.

Marko, Kurt. *The NSA and big data: what it can learn* [en línea]. Information Week, 2013 [Consulta: 10 agosto 2016]. Disponible en: <http://www.informationweek.com/big-data/big-data-analytics/the-nsa-and-big-data-what-it-can-learn/d/d-id/1110818?>.

Mayer-Schönberger, Viktor; Cukier, Kenneth. *Big data: la revolución de los datos masivos*. Madrid: Turner Publicaciones S.L., 2013. ISBN 978-84-15832-10-2.

McKinsey Global Institute. *Big Data: the next frontier for innovation, competition and productivity*. McKinsey & Company, 2011.

Microsoft Azure. *Hadoop. ¿Qué es Hadoop?* [en línea]. Microsoft Corporation [Consulta: 23 febrero 2016]. Disponible en: <https://azure.microsoft.com/es-es/solutions/hadoop/>.

Moraes, Claude. *Documento de Trabajo 1 sobre los programas de vigilancia de Estados Unidos y la UE y su repercusión sobre los derechos fundamentales europeos* [en línea]. Parlamento Europeo, 2015 [Consulta: 7 junio 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?type=COMPARL&reference=PE-524.799&format=PDF&language=ES&secondRef=01>.

NDR.de. *Snowden-Interview: transcript* [en línea]. Norddeutscher Rundfunk, 2014 [Consulta: 4 junio 2016]. Disponible en: <https://web.archive.org/web/20140128224400/http://www.ndr.de/ratgeber/netzwelt/snowden277_page-1.html>.

OBS Business School. *Big Data 2015* [en línea]. Barcelona: Universitat de Barcelona, 2015 [Consulta: 4 mayo 2016]. Disponible en: <http://www.obs-edu.com/es/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet>.

OBS Business School. *Big Data 2016* [en línea]. Barcelona: Universitat de Barcelona, 2016 [Consulta: 4 mayo 2016]. Disponible en: <http://www.obs-edu.com/es/noticias/estudio-obs/estudio-obs-big-data-2016>.

Organización de los Estados Americanos. *Declaración conjunta sobre programas de vigilancia y su impacto en la libertad de expresión* [en línea]. Washington DC: OEA, 2013 [Consulta: 20 abril 2016]. Disponible en: <http://www.oas.org/es/cidh/expresion/showarticle.asp?artID=927&>.

Paniagua, Arturo J. *Saben quién eres* [en línea]. RTVE, 2013 [Consulta: 11 febrero 2016]. Disponible en: <http://www.rtve.es/television/20131113/saben-quien-eres-torres-reyes-del-big-data/790801.shtml>.

Parlamento Europeo. *Los eurodiputados debaten sobre la protección de los datos enviados a Estados Unidos* [en línea]. Justicia y Asuntos de Interior, Relaciones Exteriores, 2016 [Consulta: 20 agosto 2016]. Disponible en: <http://www.europarl.europa.eu/news/es/news-room/20160523STO28427/los-eurodiputados-debaten-sobre-la-protecci%C3%B3n-de-los-datos-enviados-a-eeuu>.

Parlamento Europeo. *Resolución del Parlamento Europeo, de 29 de octubre de 2015, sobre el seguimiento de la Resolución del Parlamento Europeo, de 12 de marzo de 2014, relativa a la vigilancia electrónica masiva de los ciudadanos de la UE (2015/2635 RSP)* [en línea]. Estrasburgo: Parlamento Europeo, 2015. [Consulta: 4 agosto 2016]. Disponible en: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2015-0388+0+DOC+XML+V0//ES>.

Parlamento Europeo; Consejo de la Unión Europea. *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)* [en línea]. Diario Oficial de la Unión Europea, 2016 [Consulta: 4 agosto 2016]. Disponible en: <https://www.boe.es/doue/2016/119/L00001-00088.pdf>.

Penney, Jon. *Chilling effects: online surveillance and Wikipedia use* [en línea]. Oxford Internet Institute, University of Oxford, 2016 [Consulta: 15 agosto 2016]. Disponible en: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645>.

Pires, Hindenburgo Francisco. Geografia das indústrias globais de vigilancia em massa: limites à liberdade de expressão e organização na Internet. *Aracne, revista electrónica sobre Geografía y Ciencias Sociales* [en línea]. Universitat de Barcelona. Abril de 2014, núm. 183 [Consulta: 15 enero 2016]. Disponible en: <http://www.ub.edu/geocrit/aracne/aracne-183.htm>.

Poitras, Laura. *Citizenfour* [en línea]. 2014, 114 min. [Consulta: 16 enero 2016]. Disponible en: <http://peliculasio.com/citizenfour>.

Poitras, Laura; Gude, Hubert; Rosenbach, Marcel. *Mass Data: transfers from Germany Aid US Surveillance* [en línea]. Spiegel Online International, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://www.spiegel.de/international/world/german-intelligence-sends-massive-amounts-of-data-to-the-nsa-a-914821.html>.

Qlik Tech International AB. *Qlik* [en línea] 1993-2016 [Consulta: 5 febrero 2016]. <http://global.qlik.com/es>.

Quintana, Yolanda. *Todos los programas de espionaje de la NSA desvelados por Snowden* [en línea]. Eldiario.es, 2014 [Consulta: 2 febrero 2016]. Disponible en: <http://www.eldiario.es/turing/vigilancia_y_privacidad/NSA-programas-vigilancia-desvelados-Snowden_0_240426730.html>.

Reporters without Borders. *Enemies of the Internet. 2013 Report* [en línea]. París: International Secretariat Reporters without Borders, 2013 [Consulta: 16 julio 2016]. Disponible en: <http://surveillance.rsf.org/en/wp-content/uploads/sites/2/2013/03/enemies-of-the-internet_2013.pdf>.

RT. *El objetivo del 'Club de los Cinco Ojos' es "la supremacía económica sobre otros países"* [en línea]. RT, 2013 [Consulta: 15 mayo 2016] Disponible en: <https://actualidad.rt.com/actualidad/view/110072-club-cinco-ojos-supremacia-economica-espionaje>.

RTVE.es. *La NSA tiene capacidad para espiar el 75% del tráfico de Internet de Estados Unidos* [en línea]. Corporación RTVE, 2013 [Consulta: 22 mayo 2016]. Disponible en: <http://www.rtve.es/noticias/20130821/nsa-tiene-capacidad-para-espiar-75-del-trafico-internet-estados-unidos/741820.shtml>.

Salamanca Aguado, Esther. El respeto a la vida privada y a la protección de datos personales en el contexto de la vigilancia masiva de comunicaciones. *Revista del Instituto Español de Estudios Estratégicos* [en línea]. 2014, núm. 4 [Consulta: 13 enero 2016]. Disponible en: <https://dialnet.unirioja.es/servlet/articulo?codigo=4900470>.

Sanger, David; Perlroth, Nicole. *NSA breached chinese servers seen as security threat* [en línea]. The New York Times, 2014 [Consulta: 13 agosto 2016]. Disponible en: <http://www.nytimes.com/2014/03/23/world/asia/nsa-breached-chinese-servers-seen-as-spy-peril.html?partner=rss&emc=rss&smid=tw-nytimes&_r=0>.

Schneier, Bruce. *Attacking TOR: how the NSA targets users' online anonymity* [en línea]. The Guardian, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.theguardian.com/world/2013/oct/04/tor-attacks-nsa-users-online-anonymity>.

Shane, Scott. *New leaked document outlines U.S. spending on intelligence agencies* [en línea]. Nueva York: The New York Times, 2013. [Consulta: 17 mayo 2016]. Disponible en: <http://www.nytimes.com/2013/08/30/us/politics/leaked-document-outlines-us-spending-on-intelligence.html?hp&pagewanted=all&_r=0>.

Snowden Doc Search. *Boundless Informant – Describing Mission Capabilities from Metadata Records* [en línea]. Journalistic Source Protection Defence Fund, 2012 [Consulta: 2 agosto 2016]. Disponible en: <https://search.edwardsnowden.com/docs/BoundlessInformant%E2%80%93DescribingMissionCapabilitiesfromMetadataRecords2013-06-08nsadocs>.

Snowden Doc Search. *CSEC cyber threat capabilities* [en línea]. Journalistic Source Protection Defence Fund, 2011 [Consulta: 11 junio 2016]. Disponible en: <https://search.edwardsnowden.com/docs/CSECCyberThreatCapabilities2015-03-23nsadocs>.

Snowden Doc Search. *Tempora: the world's largest XKeyScore* [en línea]. Journalistic Source Protection Defence Fund, 2012 [Consulta: 10 junio 2016]. Disponible en: <https://search.edwardsnowden.com/docs/TEMPORA%E2%80%94%E2%80%9CTheWorld%E2%80%99sLargestXKEYSCORE%E2%80%9D%E2%80%94IsNowAvailabletoQualifiedNSAUsers2014-06-18nsadocs>.

Snowden Doc Search. *User's Guide for PRISM Skype Collection* [en línea]. Journalistic Source Protection Defence Fund, 2012 [Consulta: 20 julio 2016]. Disponible en: <https://search.edwardsnowden.com/docs/User%E2%80%99sGuideforPRISMSkypeCollection2014-12-28nsadocs>.

Snowden, Edward. *Introductory Statement* [en línea]. Comisión de Libertades Civiles, Justicia y Asuntos de Interior (LIBE), Parlamento Europeo, 2014 [Consulta: 13 agosto 2016]. Disponible en: <http://www.europarl.europa.eu/document/activities/cont/201403/20140307ATT80674/20140307ATT80674EN.pdf>.

Somerville, David. *NSA intelligence platforms* [en línea]. Mindmeister, 2014 [Consulta: 31 mayo 2016]. Disponible en: <https://www.mindmeister.com/es/308518551/the-national-security-agency-operates-more-than-500-separate-signals-intelligence-platforms-employs-roughly-30-000-civilians-and-military-budget-10-billion>.

Sosa Troya, María. *¿Cómo legislan los Gobiernos sobre vigilancia masiva?* [en línea]. Madrid: El País, 2015 [Consulta: 28 julio 2016]. Disponible en: <http://internacional.elpais.com/internacional/2015/06/05/actualidad/1433515872_277281.html>.

Spiegel Online. *NSA-Dokumente, so knackt der Geheimdienst Internetkonten* [en línea]. Der Spiegel, 2013. Diapositivas 13 y 14 [Consulta: 2 julio 2016]. Disponible en: <http://www.spiegel.de/fotostrecke/nsa-dokumente-so-knackt-der-geheimdienst-internetkonten-fotostrecke-105326-13.html>.

Spiegel Online. *Überwachung: BND leitet massenhaft Metadaten an die NSA weiter* [en línea]. Hamburg: Der Spiegel, 2013 [Consulta: 4 julio 2016]. Disponible en: <http://www.spiegel.de/netzwelt/netzpolitik/bnd-leitet-laut-spiegel-massenhaft-metadaten-an-die-nsa-weiter-a-914682.html>.

Sqrrl. *How to choose a NoSQL database* [en línea]. Sqrrl Enterprise, 2013 [Consulta: 3 agosto 2016]. Disponible en: <https://sqrrl.com/how-to-choose-a-nosql-database/>.

Tableau Software. *Tableau* [en línea] 2013-2016 [Consulta: 5 febrero 2016]. Disponible en: <http://www.tableau.com/es-es>.

Taigman, Yaniv [et.al]. *DeepFace: Closing the Gap to Human-Level Performance in Face Verification* [en línea]. Facebook, 2014 [Consulta: 17 abril 2016]. Disponible en: <https://www.facebook.com/publications/546316888800776/>.

Tascón, Mario. Introducción: Big Data. Pasado, presente y futuro. *Telos: Revista de Pensamiento sobre Comunicación, Tecnología y Sociedad* [en línea]. Julio-septiembre de 2013, núm.95 [Consulta: 17 junio 2016]. Disponible en: <http://telos.fundaciontelefonica.com/seccion=1268&idioma=es_ES&id=2013062110090002&activo=6.do>.

The Citizen Lab. *For Their Eyes Only: The Commercialization of Digital Spying* [en línea]. Toronto: Munk School of Global Affairs, University of Toronto, 2013 [Consulta: 15 julio 2016]. Disponible en: <https://citizenlab.org/2013/04/for-their-eyes-only-2/>.

The Citizen Lab. *Publications* [en línea]. Toronto: Munk School of Global Affairs, University of Toronto, 2013-2015 [Consulta: 15 julio 2016]. Disponible en: <https://citizenlab.org/publications/>.

The Copenhagen Post Online. *Denmark in US spy agreement?* [en línea]. The Copenhagen Post, 2013 [Consulta: 8 julio 2016]. Disponible en: <http://cphpost.dk/news/international/denmark-in-us-spy-agreement.html>.

The Guardian. *Brazil accuses Canada of spying after NSA leaks* [en línea]. The Guardian, 2013 [Consulta: 13 agosto 2016]. Disponible en: <https://www.theguardian.com/world/2013/oct/08/brazil-accuses-canada-spying-nsa-leaks>.

The Guardian. *"Tor Stinks" presentation – read the full document* [en línea]. The Guardian, 2013 [Consulta: 2 julio 2016]. Disponible en: <http://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document>.

Timberg, Craig. *NSA paying US companyies for access to communications networks* [en línea]. The Washington Post, 2013 [Consulta: 2 julio 2016]. Disponible en: <https://www.washingtonpost.com/world/national-security/nsa-paying-us-companies-for-access-to-communications-networks/2013/08/29/5641a4b6-10c2-11e3-bdf6-e4fc677d94a1_story.html>.

Timberg, Craig. *NSA slide shows surveillance of undersea cables* [en línea]. The Washington Post, 2013 [Consulta: 20 julio 2016]. Disponible en: <https://www.washingtonpost.com/business/economy/the-nsa-slide-you-havent-seen/2013/07/10/32801426-e8e6-11e2-aa9f-c03a72e2d342_story.html>.

TRC. *Conceptos básicos de Big Data* [en línea]. Madrid: TRC, 2014 [Consulta: 28 marzo 2016] Disponible en: <http://www.trc.es/pdf/descargas/big_data.pdf>.

Tremlett, Giles. *US offers to spy on ETA for Spain* [en línea]. The Guardian, 2001 [Consulta: 9 julio 2016]. Disponible en: <https://www.theguardian.com/world/2001/jun/15/spain.usa>.

Tsukayama, Hayley. *PalTalk: The Prism company that you've never heard of* [en línea]. The Washington Post, 2013 [Consulta: 24 febrero 2016]. Disponible en: <https://www.washingtonpost.com/business/technology/paltalk-the-prism-company-that-youve-never-heard-of/2013/06/07/02a0f2c4-cf79-11e2-8f6b-67f40e176f03_story.html>.

Velasco, JJ. *La NSA espió a Naciones Unidas interceptando su sistema de videoconferencia* [en línea]. Hipertextual, 2013 [Consulta: 14 agosto 2016]. Disponible en: <https://hipertextual.com/2013/08/nsa-espio-a-naciones-unidas>.

WikiLeaks. *NSA targets world leaders for U.S. geopolitical interests* [en línea]. WikiLeaks, 2016 [Consulta: 14 agosto 2016]. Disponible en: <https://wikileaks.org/nsa-201602/>.

WikiLeaks. *Public Library of U.S. Diplomacy* [en línea]. WikiLeaks, 2010 [Consulta: 14 agosto 2016]. Disponible en: <https://wikileaks.org/plusd/cables/09STATE80163_a.html#efmJZLJeM>.

WikiLeaks. *Target Tokyo* [en línea]. WikiLeaks, 2015 [Consulta: 13 agosto 2016]. Disponible en: <https://wikileaks.org/nsa-japan/>.

WikiLeaks. *The spy files* [en línea]. WikiLeaks, 2011 [Consulta: 15 julio 2016]. Disponible en: <https://wikileaks.org/the-spyfiles.html>.

Yárnoz, Carlos. *Francia aprueba la ley que permite espiar sin control judicial* [en línea]. El País, 2015 [Consulta: 11 agosto 2016]. Disponible en: <http://internacional.elpais.com/internacional/2015/05/05/actualidad/1430823773_342509.html>.

Zamora, Inma. *Big Data: ¿vidas privadas al alcance de todos?* [en línea]. Madrid: ABC, 2013 [Consulta: 23 febrero 2016]. Disponible en: <http://www.abc.es/tecnologia/informatica-software/20131028/abci-entrevista-data-201310221252.html>.

Zetter, Kim. *How to detect sneaky NSA "Quantum Insert" attacks* [en línea]. The Wired, 2015 [Consulta: 2 julio 2016]. Disponible en: <https://www.wired.com/2015/04/researchers-uncover-method-detect-nsa-quantum-insert-hacks/>.